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? (3 marks)

Answer : Below are the observation from Categorical Variables from Dataset:

I. **Season**: Almost 32% of the bike booking were happening in fall with a median of over 5000 booking (for the period of 2 years). This was followed by Summer & Winter with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

II. **mnth**: Almost 10% of the bike booking were happening in the months May, June, Jul, Aug & Sep with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

III. **weathersit**: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

IV. **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday can't be a good predictor for the dependent variable.

V. **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

VI. **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

In Our Final modal also we saw that winter and summer is impacting the model as below:

I. **Summer**: A coefficient value of ‘0.0881’ indicated that w.r.t Winter, a unit increase in Summer variable increases the bike hire numbers by 0.0881 units.

II. **Winter**: A coefficient value of ‘0.1293’ indicated that w.r.t Summer, a unit increase in Winter variable Increases the bike hire numbers by 0.1293 units.

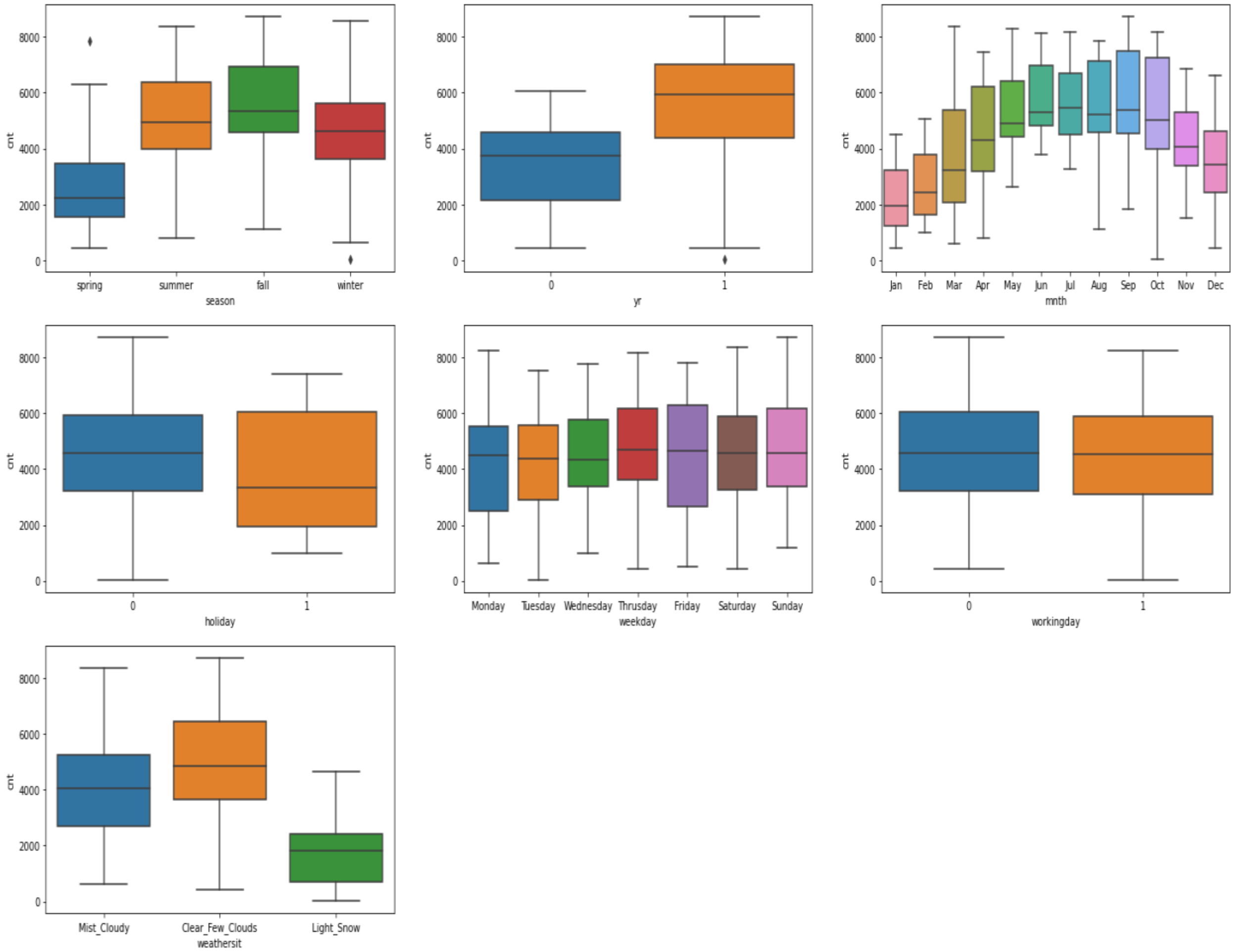
III. **yr**: A coefficient value of ‘0.2329’ indicated that a unit increase in yr variable, increases the bike hire numbers by 0.2329 units.

IV. **Sep**: A coefficient value of ‘0.1012’ , a unit increase in Sep variable increases the bike hire numbers by 0.1012 units.

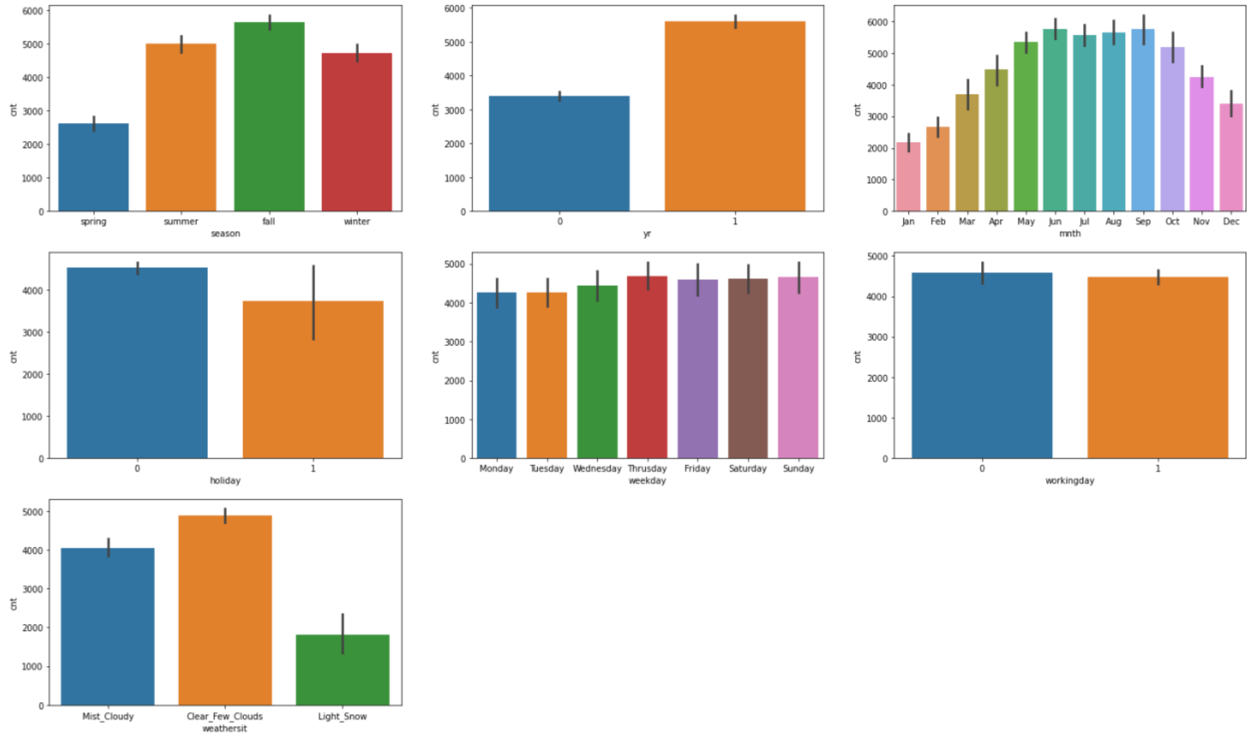
V. **holiday**: A coefficient value of ‘-0.0987’ , a unit increase in holiday variable decrease the bike hire numbers by 0.0987 units.

Please see the screenshot below:

Box Plot



Bar-PLot



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

Answer:

Dop_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.

If we don't use "drop_first" we will get a redundant feature.

Example:

If we have a feature "Is_male", we use "get_dummies" we will get two features "Is_male_0" and "Is_male_1", but if you look carefully they are redundant actually. we just need one of them, the other one will the exact opposite of the other.

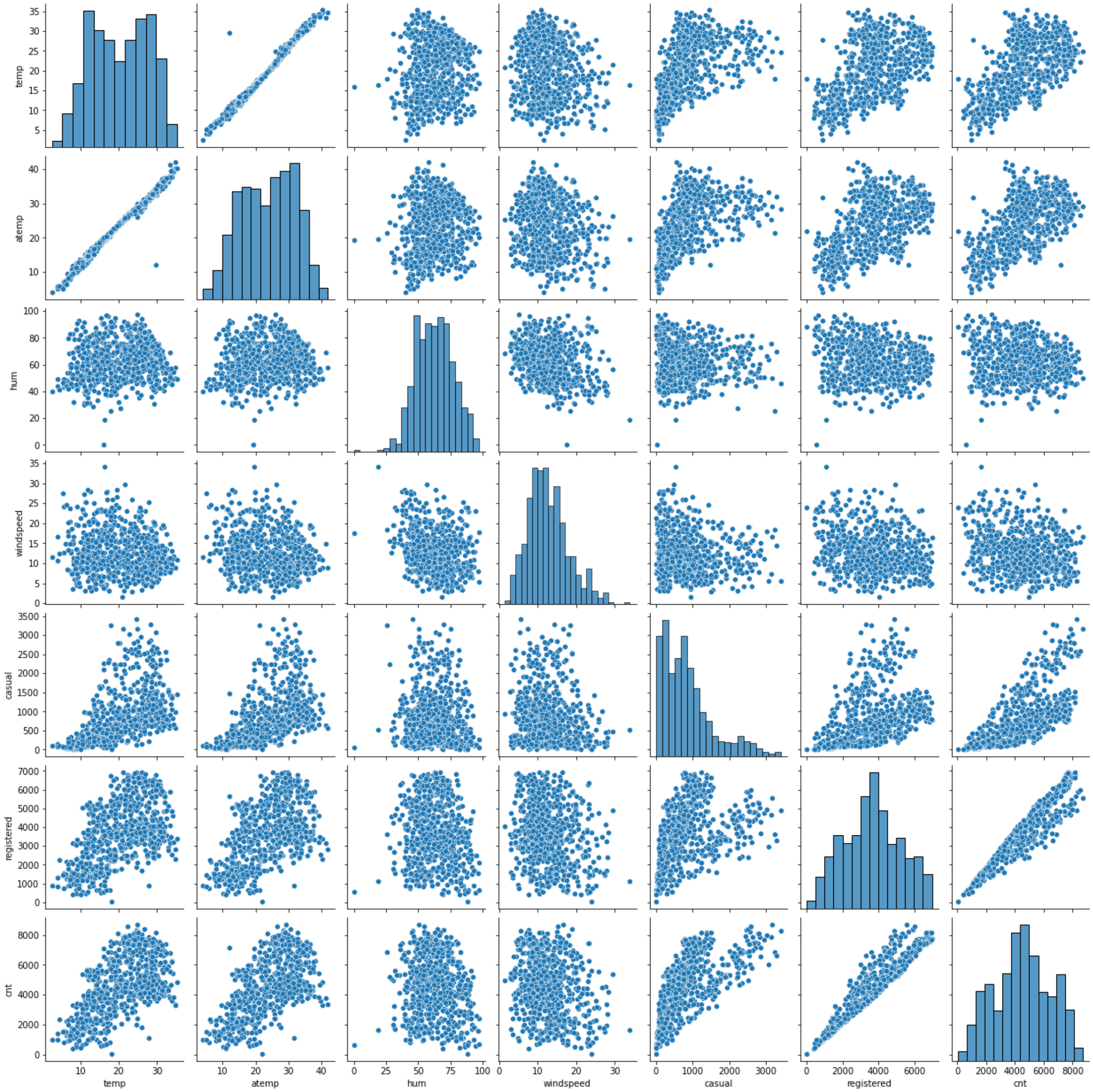
3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

Below are the observation of pair plot among the numerical variables:

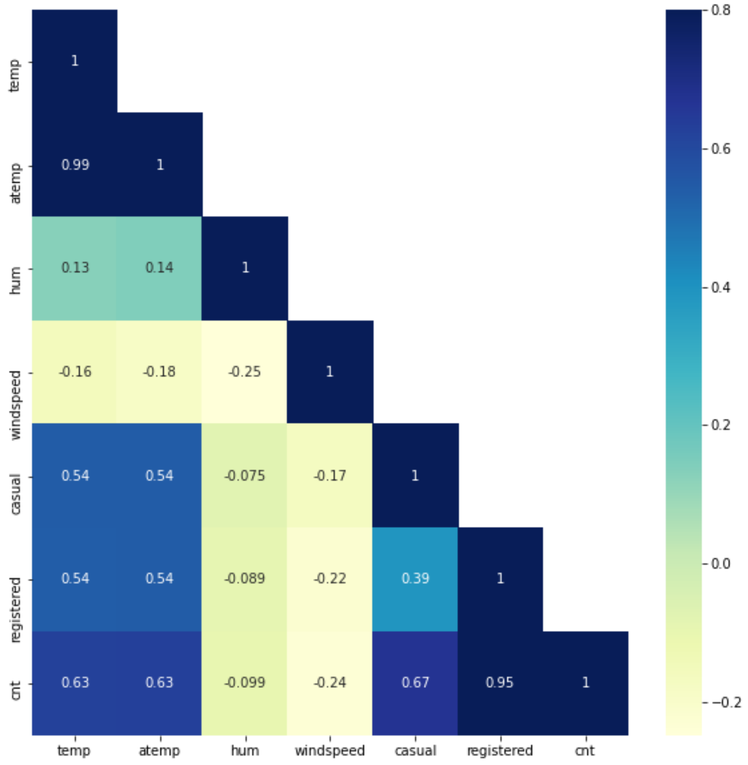
I. There is a linear relation between 'temp' and 'atemp' variable with the predictor 'cnt'.

Pair Plot



II. The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable.

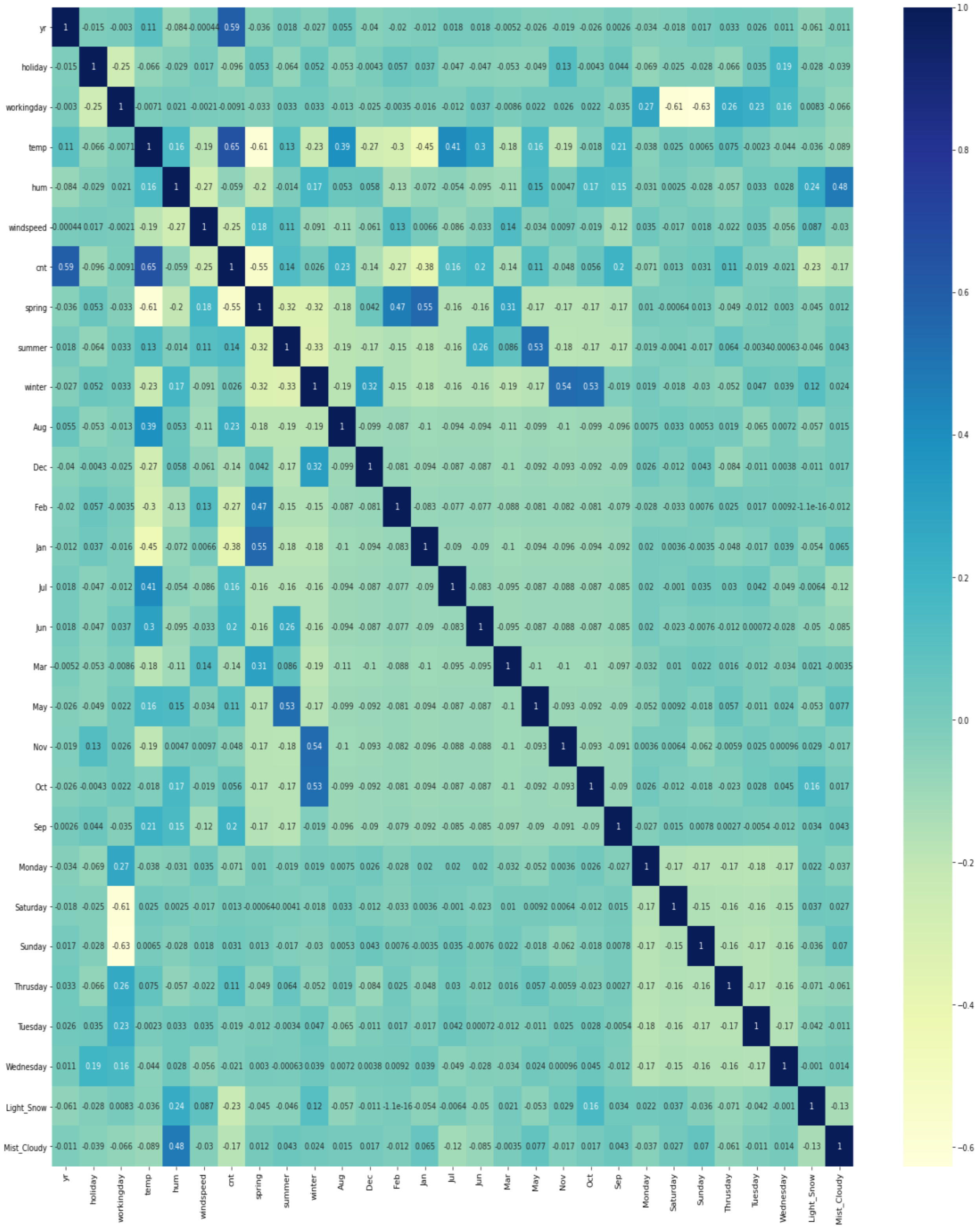
Heatmap



III. INSIGHTS

a. from above corr matrix we can say that temp and atemp have a strong relationship means it will create multicollinearity issue
b. cnt and registered is also strongly correlated
c. windspeed and cnt are negatively related which means during windspeed people don't prefer to rent a bike
d. cnt and casual are also strongly related
e. cnt and registered are also strongly related

After Creating Dummy Variable Please see the correlation(Screenshot Below):



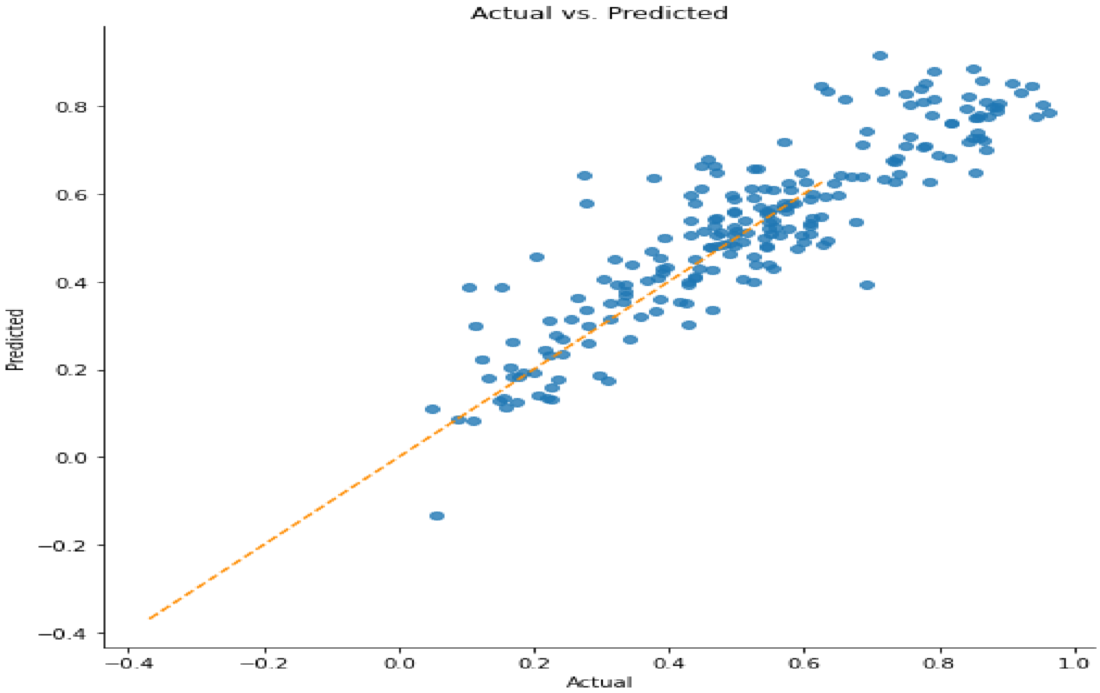
	yr	holiday	workingday	temp	hum	windspeed	cnt	spring	summer	winter	Aug	Dec	Feb	Jan	Jul	Jun	Mar	May	Nov	Oct	Sep	Monday	Saturday	Sunday	Thrusday	Tuesday	Wednesday	Light_Snow	Mist_Cloudy
yr	1	-0.015	-0.003	0.11	-0.084	-0.00044	0.59	-0.036	0.018	-0.027	0.055	-0.04	-0.02	-0.012	0.018	0.018	-0.0052	-0.026	-0.019	-0.026	0.0026	-0.034	-0.018	0.017	0.033	0.026	0.011	-0.061	-0.011
holiday	-0.015	1	-0.25	-0.066	-0.029	0.017	-0.096	0.053	-0.064	0.052	-0.053	-0.0043	0.057	0.037	-0.047	-0.047	-0.053	-0.049	0.13	-0.0043	0.044	-0.069	-0.025	-0.028	-0.066	0.035	0.19	-0.028	-0.039
workingday	-0.003	-0.25	1	-0.0071	0.021	-0.0021	-0.0091	-0.033	0.033	0.033	-0.013	-0.025	-0.0035	-0.016	-0.012	0.037	-0.0086	0.022	0.026	0.022	-0.035	0.27	-0.61	-0.63	0.26	0.23	0.16	0.0083	-0.066
temp	0.11	-0.066	-0.0071	1	0.16	-0.19	0.65	-0.61	0.13	-0.23	0.39	-0.27	-0.3	-0.45	0.41	0.3	-0.18	0.16	-0.19	-0.018	0.21	-0.038	0.025	0.0065	0.075	-0.0023	-0.044	-0.036	-0.089
hum	-0.084	-0.029	0.021	0.16	1	-0.27	-0.059	-0.2	-0.014	0.17	0.053	0.058	-0.13	-0.072	-0.054	-0.095	-0.11	0.15	0.0047	0.17	0.15	-0.031	0.0025	-0.028	-0.057	0.033	0.028	0.24	0.48
windspeed	-0.00044	0.017	-0.0021	-0.19	-0.27	1	-0.25	0.18	0.11	-0.091	-0.11	-0.061	0.13	0.0066	-0.086	-0.033	0.14	-0.034	0.0097	-0.019	-0.12	0.035	-0.017	0.018	-0.022	0.035	-0.056	0.087	-0.03
cnt	0.59	-0.096	-0.0091	0.65	-0.059	-0.25	1	-0.55	0.14	0.026	0.23	-0.14	-0.27	-0.38	0.16	0.2	-0.14	0.11	-0.048	0.056	0.2	-0.071	0.013	0.031	0.11	-0.019	-0.021	-0.23	-0.17
spring	-0.036	0.053	-0.033	-0.61	-0.2	0.18	-0.55	1	-0.32	-0.32	-0.18	0.042	0.47	0.55	-0.16	-0.16	0.31	-0.17	-0.17	-0.17	-0.17	0.01	-0.00064	0.013	-0.049	-0.012	0.003	-0.045	0.012
summer	0.018	-0.064	0.033	0.13	-0.014	0.11	0.14	-0.32	1	-0.33	-0.19	-0.17	-0.15	-0.18	-0.16	0.26	0.086	0.53	-0.18	-0.17	-0.17	-0.019	-0.0041	-0.017	0.064	-0.0034	-0.00063	-0.046	0.043
winter	-0.027	0.052	0.033	-0.23	0.17	-0.091	0.026	-0.32	-0.33	1	-0.19	0.32	-0.15	-0.18	-0.16	-0.16	-0.19	-0.17	0.54	0.53	-0.019	0.019	-0.018	-0.03	-0.052	0.047	0.039	0.12	0.024
Aug	0.055	-0.053	-0.013	0.39	0.053	-0.11	0.23	-0.18	-0.19	-0.19	1	-0.099	-0.087	-0.1	-0.094	-0.094	-0.11	-0.099	-0.1	-0.099	-0.096	0.0075	0.033	0.0053	0.019	-0.065	0.0072	-0.057	0.015
Dec	-0.04	-0.0043	-0.025	-0.27	0.058	-0.061	-0.14	0.042	-0.17	0.32	-0.099	1	-0.081	-0.094	-0.087	-0.087	-0.1	-0.092	-0.093	-0.092	-0.09	0.026	-0.012	0.043	-0.084	-0.011	0.0038	-0.011	0.017
Feb	-0.02	0.057	-0.0035	-0.3	-0.13	0.13	-0.27	0.47	-0.15	-0.15	-0.087	-0.081	1	-0.083	-0.077	-0.077	-0.088	-0.081	-0.082	-0.081	-0.079	-0.028	-0.033	0.0076	0.025	0.017	0.0092	-1.1e-16	-0.012
Jan	-0.012	0.037	-0.016	-0.45	-0.072	0.0066	-0.38	0.55	-0.18	-0.18	-0.1	-0.094	-0.083	1	-0.09	-0.09	-0.1	-0.094	-0.096	-0.094	-0.092	0.02	0.0036	-0.0035	-0.048	-0.017	0.039	-0.054	0.065
Jul	0.018	-0.047	-0.012	0.41	-0.054	-0.086	0.16	-0.16	-0.16	-0.16	-0.094	-0.087	-0.077	-0.09	1	-0.083	-0.095	-0.087	-0.088	-0.087	-0.085	0.02	-0.001	0.035	0.03	0.042	-0.049	-0.0064	-0.12
Jun	0.018	-0.047	0.037	0.3	-0.095	-0.033	0.2	-0.16	0.26	-0.16	-0.094	-0.087	-0.077	-0.09	-0.083	1	-0.095	-0.087	-0.088	-0.087	-0.085	0.02	-0.023	-0.0076	-0.012	0.00072	-0.028	-0.05	-0.085
Mar	-0.0052	-0.053	-0.0086	-0.18	-0.11	0.14	-0.14	0.31	0.086	-0.19	-0.11	-0.1	-0.088	-0.1	-0.095	-0.095	1	-0.1	-0.1	-0.1	-0.097	-0.032	0.01	0.022	0.016	-0.012	-0.034	0.021	-0.0035
May	-0.026	-0.049	0.022	0.16	0.15	-0.034	0.11	-0.17	0.53	-0.17	-0.099	-0.092	-0.081	-0.094	-0.087	-0.087	-0.1	1	-0.093	-0.092	-0.09	-0.052	0.0092	-0.018	0.057	-0.011	0.024	-0.053	0.077
Nov	-0.019	0.13	0.026	-0.19	0.0047	0.0097	-0.048	-0.17	-0.18	0.54	-0.1	-0.093	-0.082	-0.096	-0.088	-0.088	-0.1	-0.093	1	-0.093	-0.091	0.0036	0.0064	-0.062	-0.0059	0.025	0.00096	0.029	-0.017
Oct	-0.026	-0.0043	0.022	-0.018	0.17	-0.019	0.056	-0.17	-0.17	0.53	-0.099	-0.092	-0.081	-0.094	-0.087	-0.087	-0.1	-0.092	-0.093	1	-0.09	0.026	-0.012	-0.018	-0.023	0.028	0.045	0.16	0.017
Sep	0.0026	0.044	-0.035	0.21	0.15	-0.12	0.2	-0.17	-0.17	-0.019	-0.096	-0.09	-0.079	-0.092	-0.085	-0.085	-0.097	-0.09	-0.091	-0.09	1	-0.027	0.015	0.0078	0.0027	-0.0054	-0.012	0.034	0.043
Monday	-0.034	-0.069	0.27	-0.038	-0.031	0.035	-0.071	0.01	-0.019	0.019	0.0075	0.026	-0.028	0.02	0.02	0.02	-0.032	-0.052	0.0036	0.026	-0.027	1	-0.17	-0.17	-0.17	-0.18	-0.17	0.022	-0.037
Saturday	-0.018	-0.025	-0.61	0.025	0.0025	-0.017	0.013	-0.00064	-0.0041	-0.018	0.033	-0.012	-0.033	0.0036	-0.001	-0.023	0.01	0.0092	0.0064	-0.012	0.015	-0.17	1	-0.15	-0.16	-0.16	-0.15	0.037	0.027
Sunday	0.017	-0.028	-0.63	0.0065	-0.028	0.018	0.031	0.013	-0.017	-0.03	0.0053	0.043	0.0076	-0.0035	0.035	-0.0076	0.022	-0.018	-0.062	-0.018	0.0078	-0.17	-0.15	1	-0.16	-0.17	-0.16	-0.036	0.07
Thrusday	0.033	-0.066	0.26	0.075	-0.057	-0.022	0.11	-0.049	0.064	-0.052	0.019	-0.084	0.025	-0.048	0.03	-0.012	0.016	0.057	-0.0059	-0.023	0.0027	-0.17	-0.16	-0.16	1	-0.17	-0.16	-0.071	-0.061
Tuesday	0.026	0.035	0.23	-0.0023	0.033	0.035	-0.019	-0.012	-0.0034	0.047	-0.065	-0.011	0.017	-0.017	0.042	0.00072	-0.012	-0.011	0.025	0.028	-0.0054	-0.18	-0.16	-0.17	-0.17	1	-0.17	-0.042	-0.011
Wednesday	0.011	0.19	0.16	-0.044	0.028	-0.056	-0.021	0.003	-0.00063	0.039	0.0072	0.0038	0.0092	0.039	-0.049	-0.028	-0.034	0.024	0.00096	0.045	-0.012	-0.17	-0.15	-0.16	-0.16	-0.17	1	-0.001	0.014
Light_Snow	-0.061	-0.028	0.0083	-0.036	0.24	0.087	-0.23	-0.045	-0.046	0.12	-0.057	-0.011	-1.1e-16	-0.054	-0.0064	-0.05	0.021	-0.053	0.029	0.16	0.034	0.022	0.037	-0.036	-0.071	-0.042	-0.001	1	-0.13
Mist_Cloudy	-0.011	-0.039	-0.066	-0.089	0.48	-0.03	-0.17	0.012	0.043	0.024	0.015	0.017	-0.012	0.065	-0.12	-0.085	-0.0035	0.077	-0.017	0.017	0.043	-0.037	0.027	0.07	-0.061	-0.011	0.014	-0.13	1

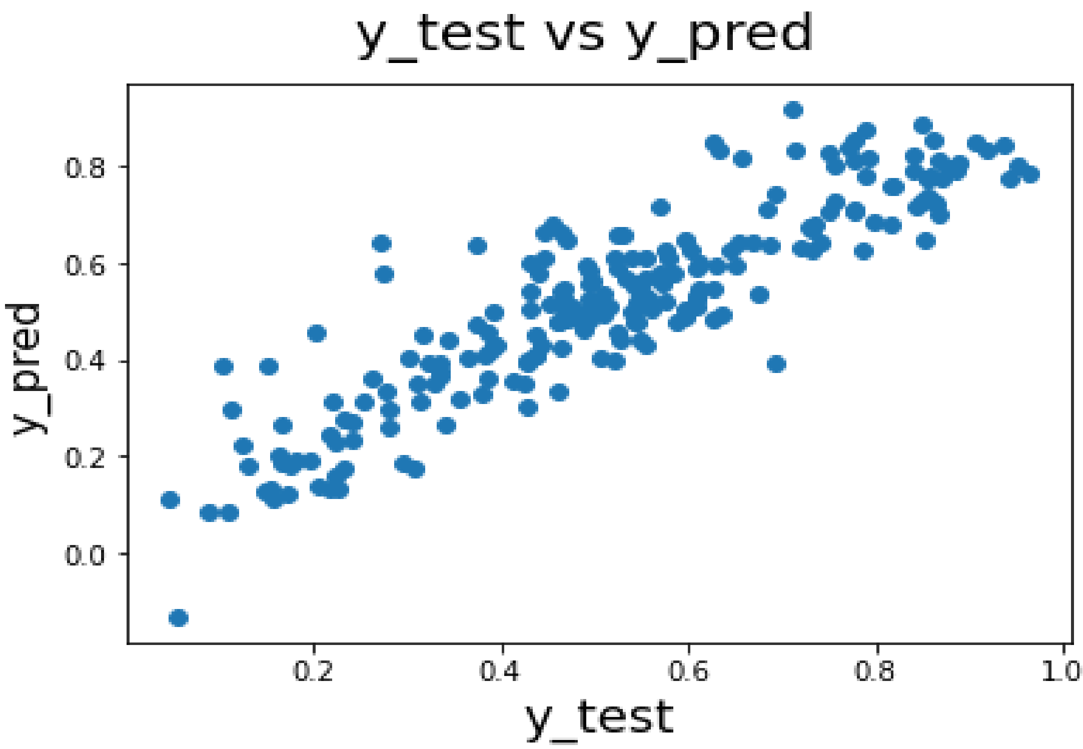
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

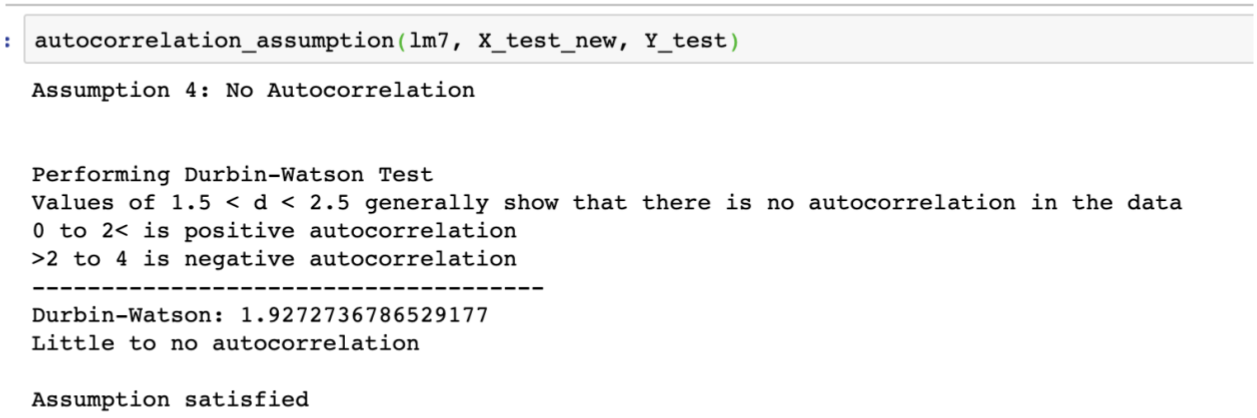
Following are the points can be considered while validate the assumption

I. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X¹ is constant, regardless of the value of X¹. An additive relationship suggests that the effect of X¹ on Y is independent of other variables. In our model 'cnt' is dependent variable and others are independent variable.





II. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

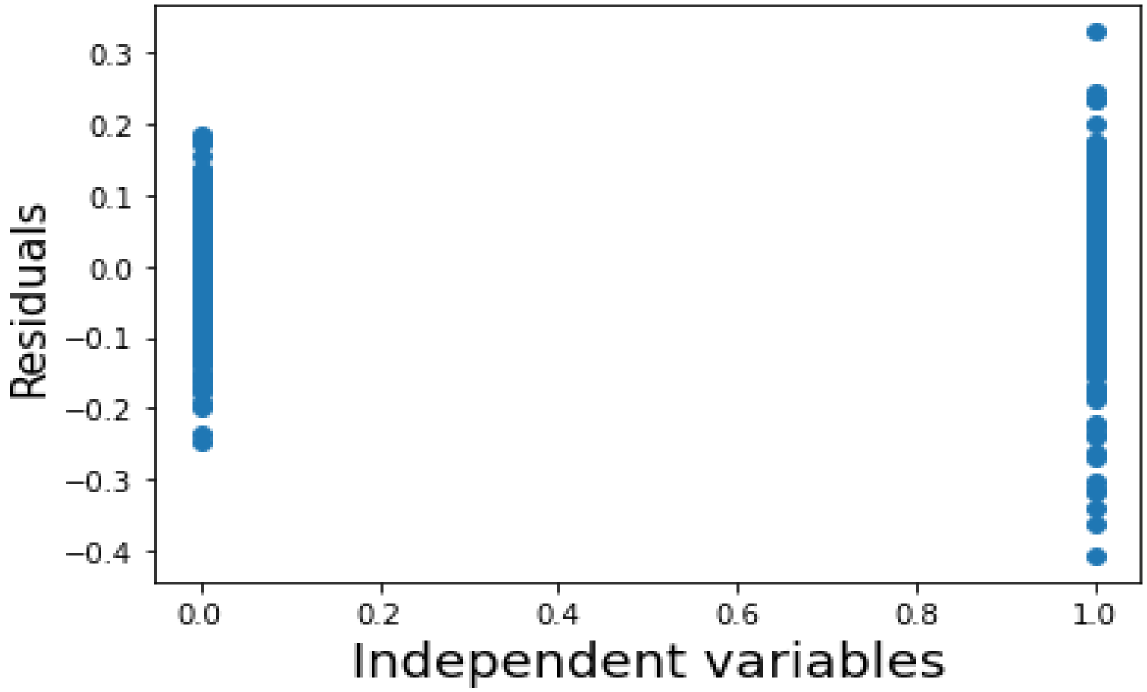


```
autocorrelation_assumption(lm7, X_test_new, Y_test)
```

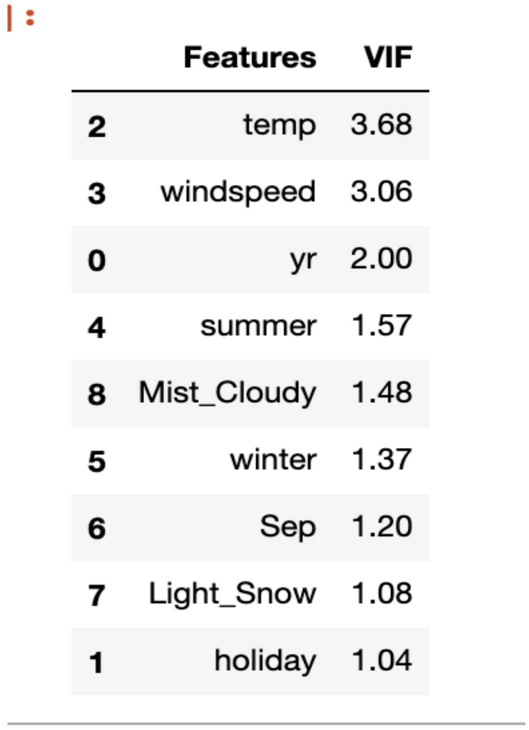
```
Assumption 4: No Autocorrelation 


Performing Durbin-Watson Test
Values of 1.5 < d < 2.5 generally show that there is no autocorrelation in the data
0 to 2< is positive autocorrelation
>2 to 4 is negative autocorrelation
-------------------------------------
Durbin-Watson: 1.9272736786529177
Little to no autocorrelation 

Assumption satisfied
```

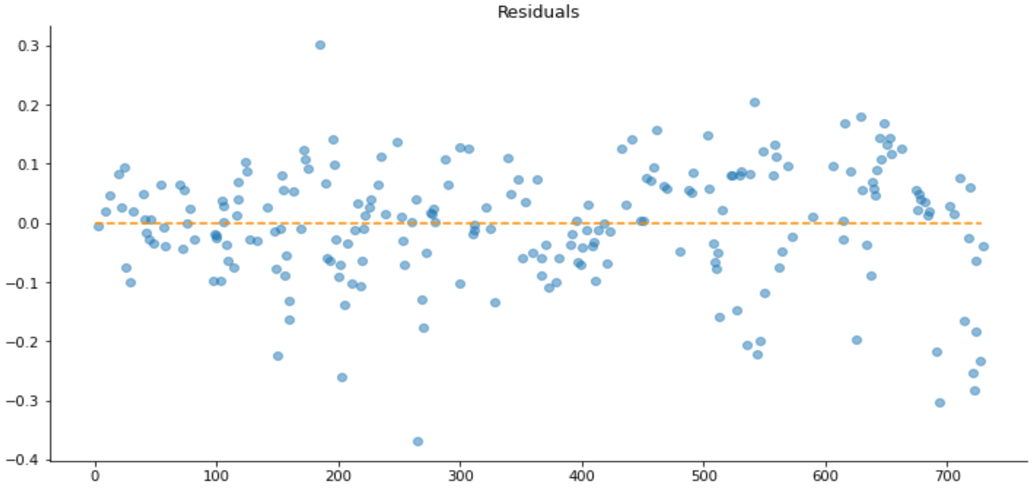


III. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

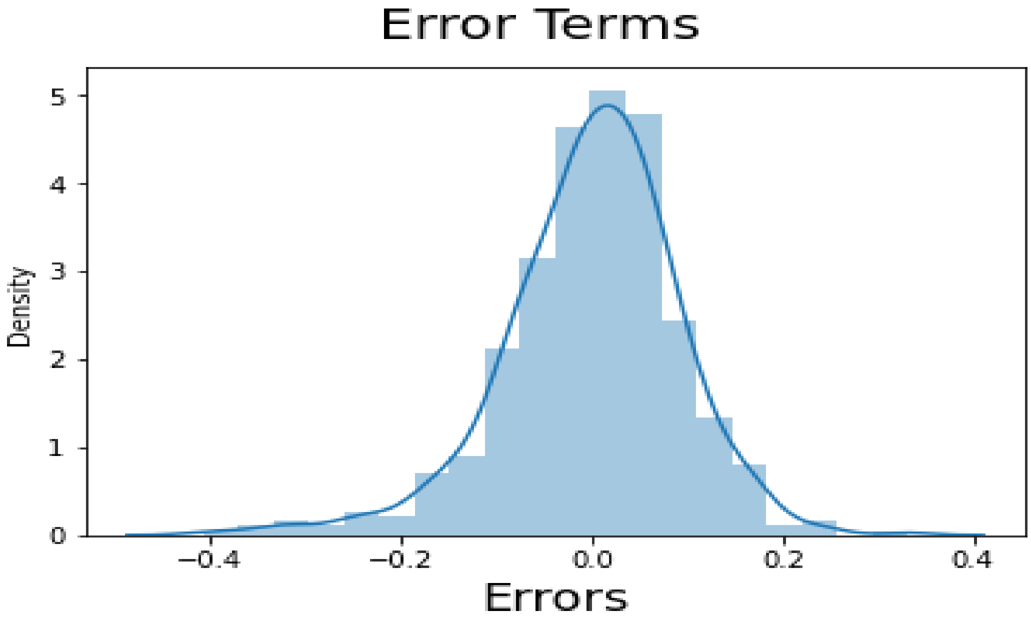

]:

	Features	VIF
2	temp	3.68
3	windspeed	3.06
0	yr	2.00
4	summer	1.57
8	Mist_Cloudy	1.48
5	winter	1.37
6	Sep	1.20
7	Light_Snow	1.08
1	holiday	1.04

IV. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.



V. The error terms must be normally distributed.(Screenshot)



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

Answer:

As per our final Model, the top 3 predictor variables that influences the bike booking are:

I. **Temperature (temp)** - A coefficient value of '0.5480' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5480 units.

II. **Light_Snow** - A coefficient value of '-0.2829' indicated that, w.r.t Mist_Cloudy, a unit increase in Light_Snow variable decreases the bike hire numbers by 0.2829 units.

III. **Year (yr)** - A coefficient value of '0.2329' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2329 units.

SO IT IS RECOMMENDED TO GIVE THESE VARIABLES UTMOST IMPORTANCE WHILE PLANNING, TO ACHIEVE MAXIMUM BOOKING.

The next best features that can also be considered are :

IV. **winter** - A coefficient value of '0.1293' indicated that w.r.t Summer, a unit increase in winter variable increases the bike hire numbers by 0.1293 units.

V. **windspeed** - A coefficient value of '-0.1532' indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1532 units.

General Subjective Questions

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

Answer :

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

Y=mX+b

Here,

Y : Dependent variable i.e. we are trying to predict.

X: Independent variable i.e. we are using to make predictions.

m: Slop of the regression line which represents the effect X has on Y

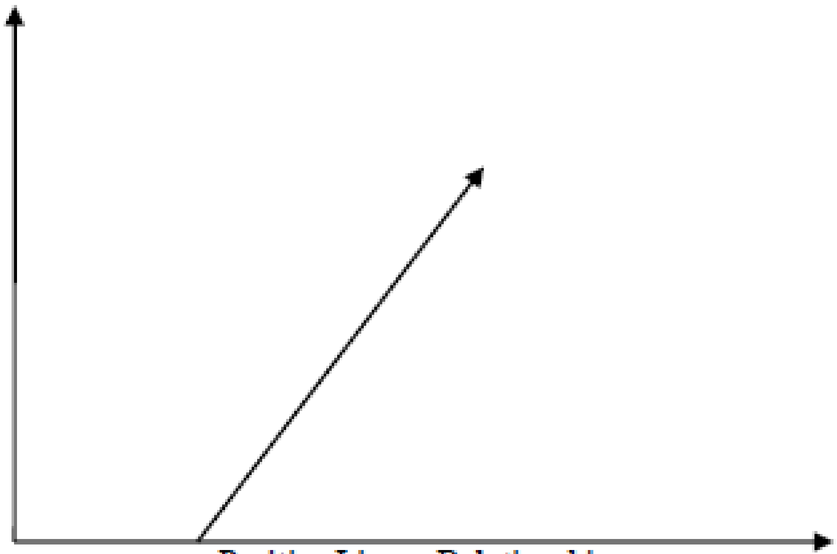
b: Constant, known as the YY-intercept.

If X = 0,Y would be equal to bb.

Furthermore, the linear relationship can be positive or negative in nature as explained below:

Positive Linear Relationship

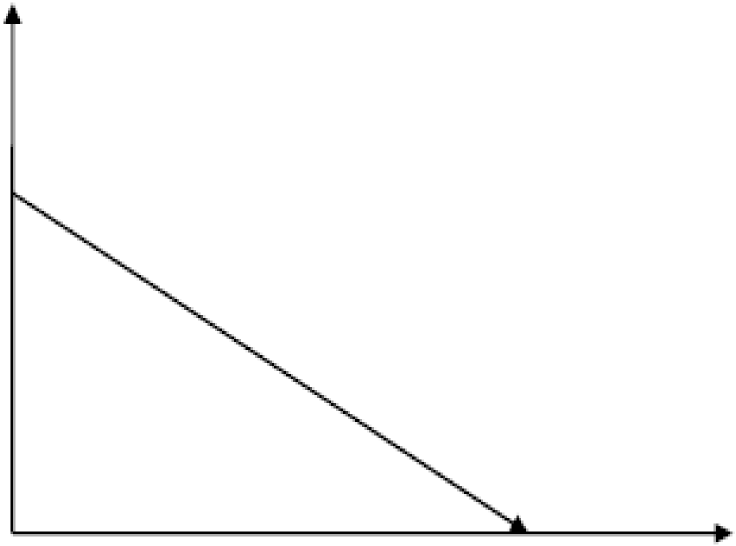
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

Negative Linear relationship:

A linear relationship will be called Negative if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Types of Linear Regression

Linear regression is of the following two types −

I. Simple Linear Regression

II. Multiple Linear Regression

Assumptions

The following are some assumptions about dataset that is made by Linear Regression model −

I. **The assumption about the form of the model:**

   It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.

II. **Assumptions about the residuals:**

   a) **Normality assumption**: It is assumed that the error terms, Œµ(i), are normally distributed.

   b) **Zero mean assumption**: It is assumed that the residuals have a mean value of zero.

   c) **Constant variance assumption**: It is assumed that the residual terms have the same (but unknown) variance, (Sigma)^2 This assumption is also known as the assumption of homogeneity or homoscedasticity.

   d) **Independent error assumption**: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

III. **Assumptions about the estimators:**

   a) The independent variables are measured without error.

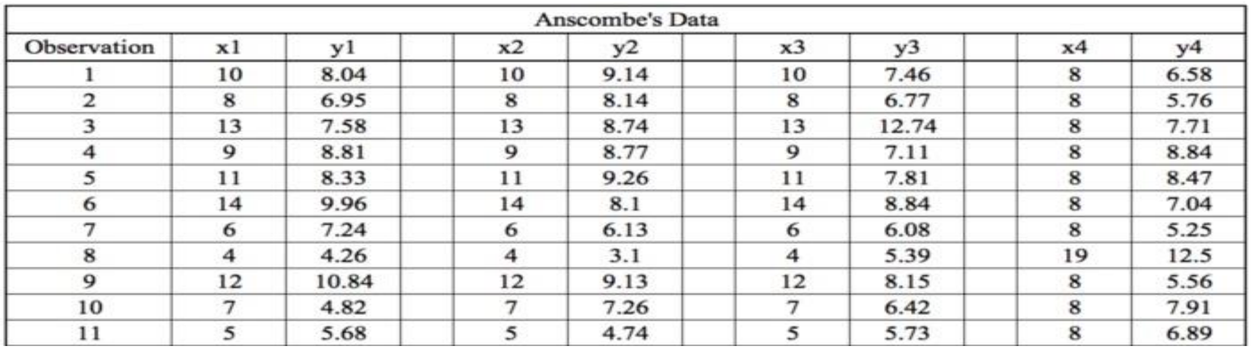
   b) The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

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

Answer:

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. 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 visualizing 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.

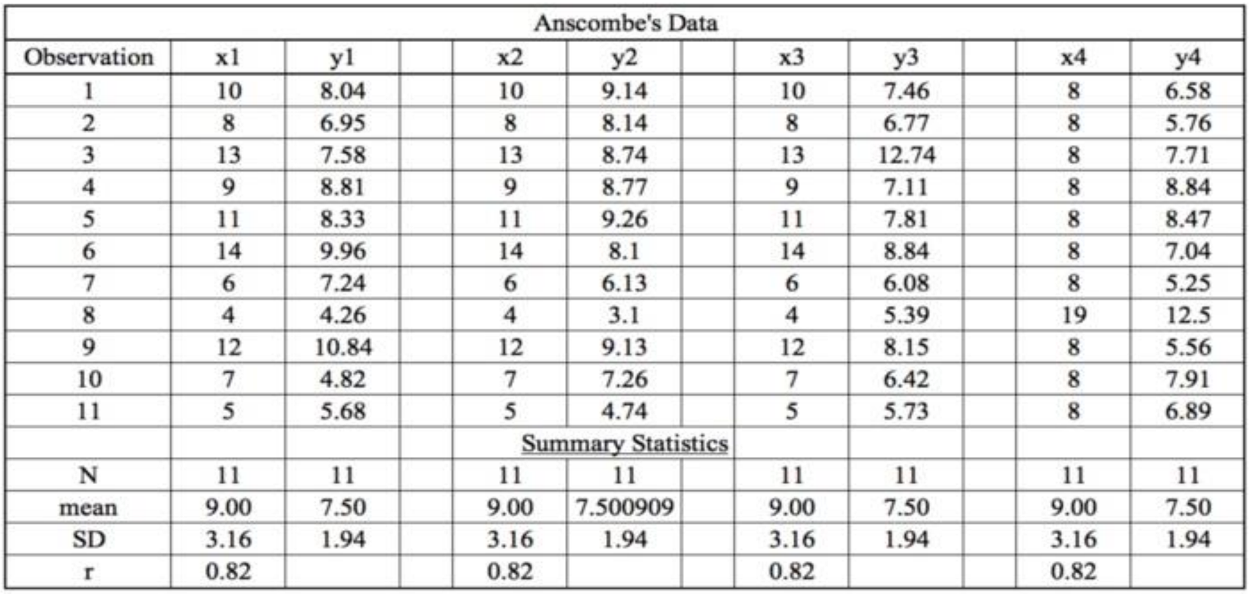
These four plots can be defined as follows:



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

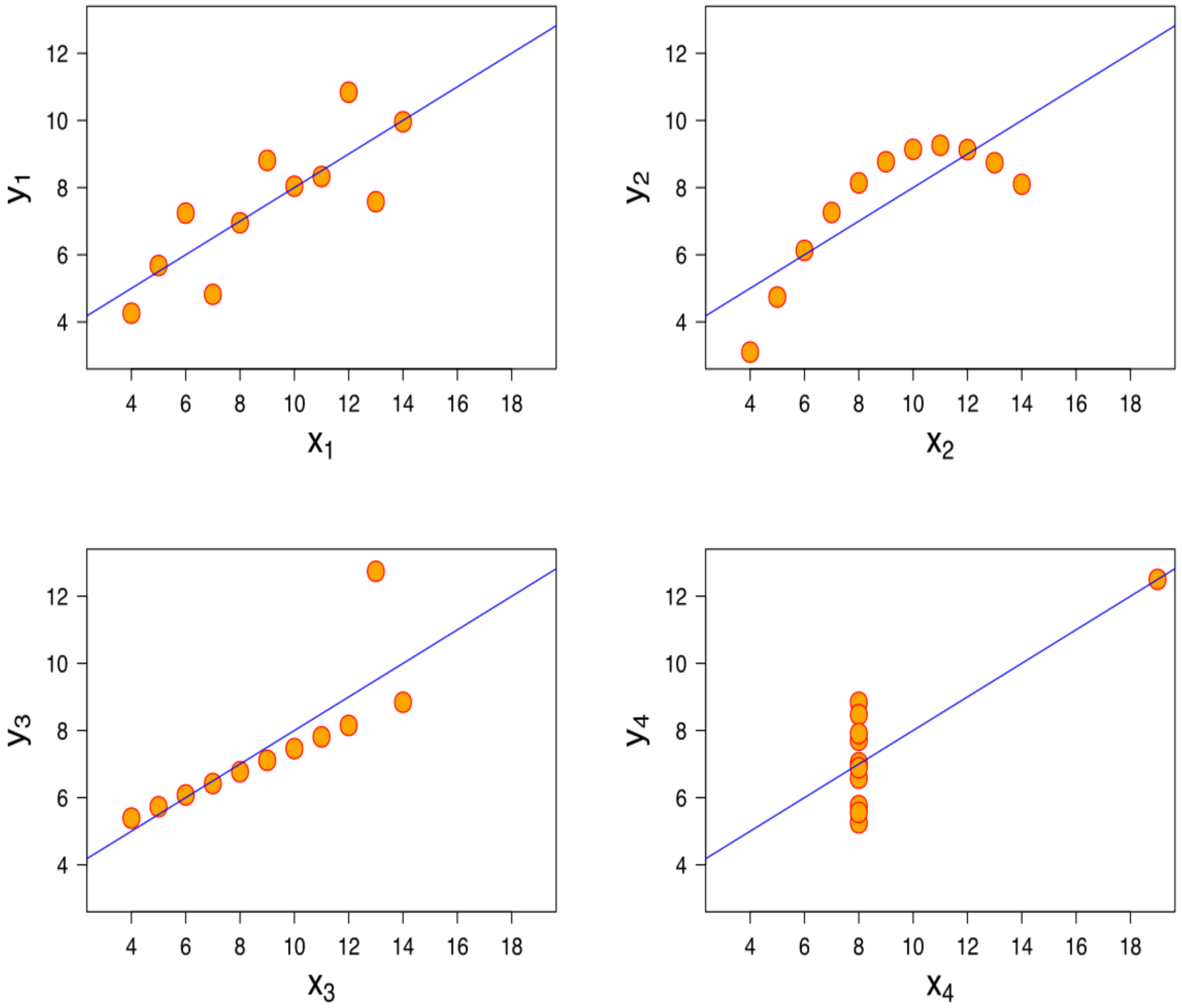
The statistical information for all these four datasets are approximately similar and can be computed as follows



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



I. **Dataset 1**: this fits the linear regression model pretty well.

II. **Dataset 2**: this could not fit linear regression model on the data quite well as the data is non-linear.

III. **Dataset 3**: shows the outliers involved in the dataset which cannot be handled by linear regression model.

IV. **Dataset 4**: shows the outliers involved in the dataset which cannot be handled by linear regression model.

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

Answer:

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

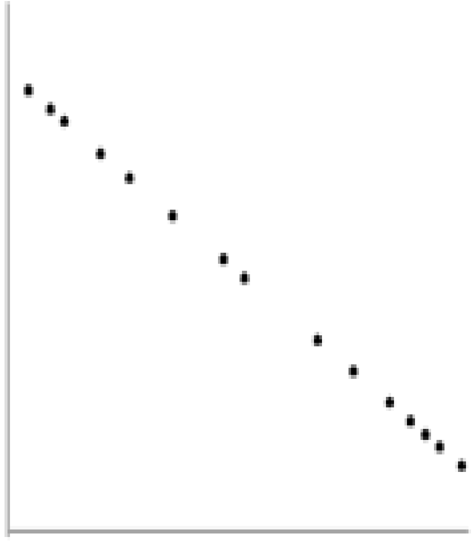
The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient

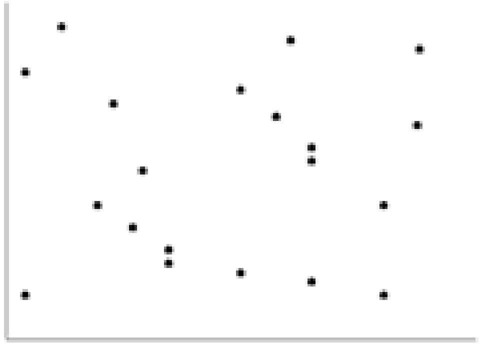
Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

R=-1



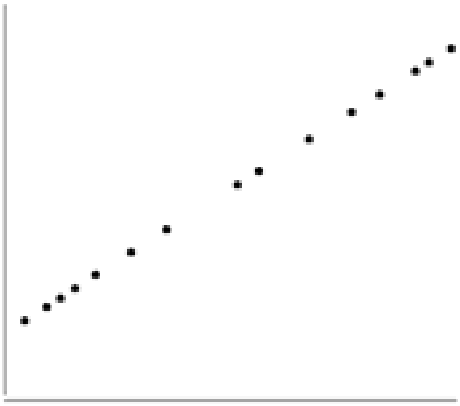
Data lie on a perfect straight line with a negative slope.

R=0



No linear relationship between the variables

R=1



Data lie on a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling 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:

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:

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

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

I. When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

II. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1

III. If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization:

Standardization is another scaling technique 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:

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

μ 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.

Difference Between Normalization and Standardization:

I. Normalization is good to use when we know that the distribution of our data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

II. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

The user has to select the variables to be included by ticking off the corresponding check boxes. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables .

If VIF is large and multicollinearity affects your analysis results, then we need to take some corrective actions before we can use multiple regression. Here are the various options:

a) One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.

b) A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.

c) The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.

d) The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

e) Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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.

Few advantages:

I. It can be used with sample sizes also

II. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

a) If two data sets

   I. come from populations with a common distribution

   II. have common location and scale

   III. have similar distributional shapes

   IV. have similar tail behavior

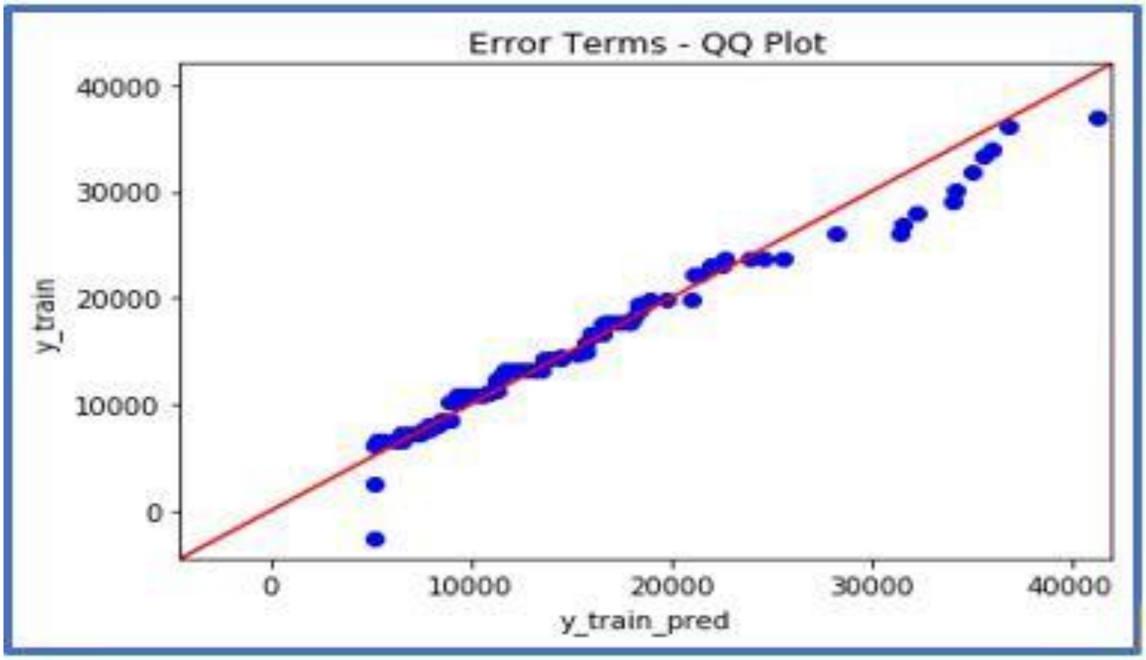
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

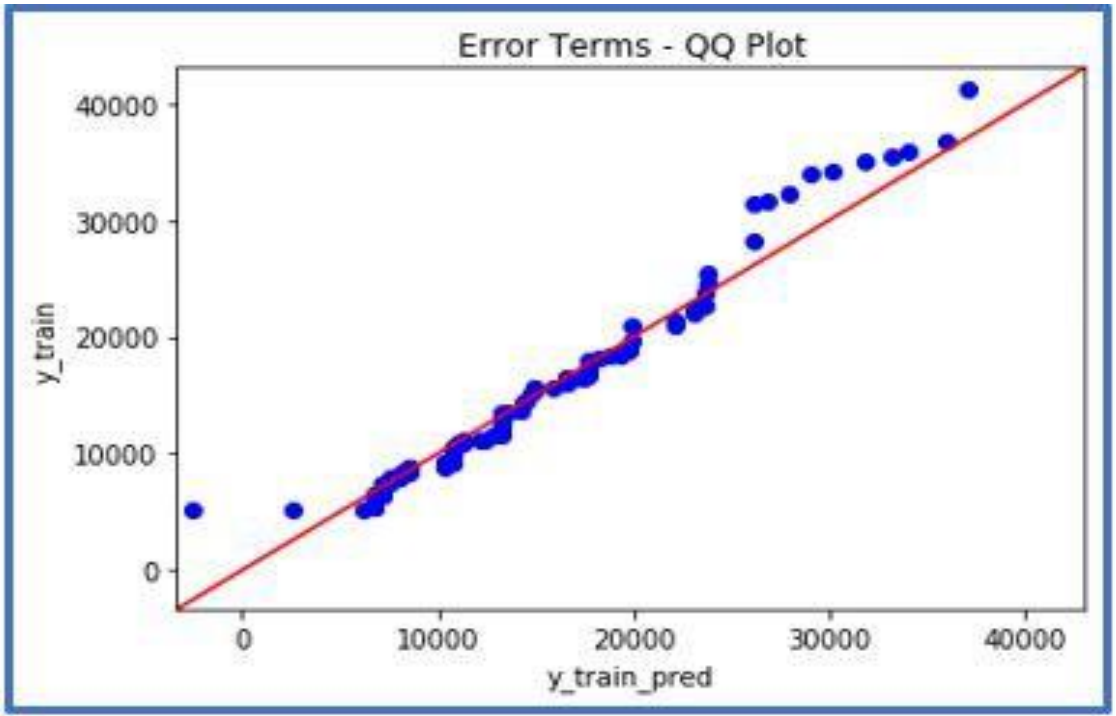
b) Below are the possible interpretations for two data sets.

I. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

II. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



III. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



c) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis