Linear Regression Bike Sharing Assignment

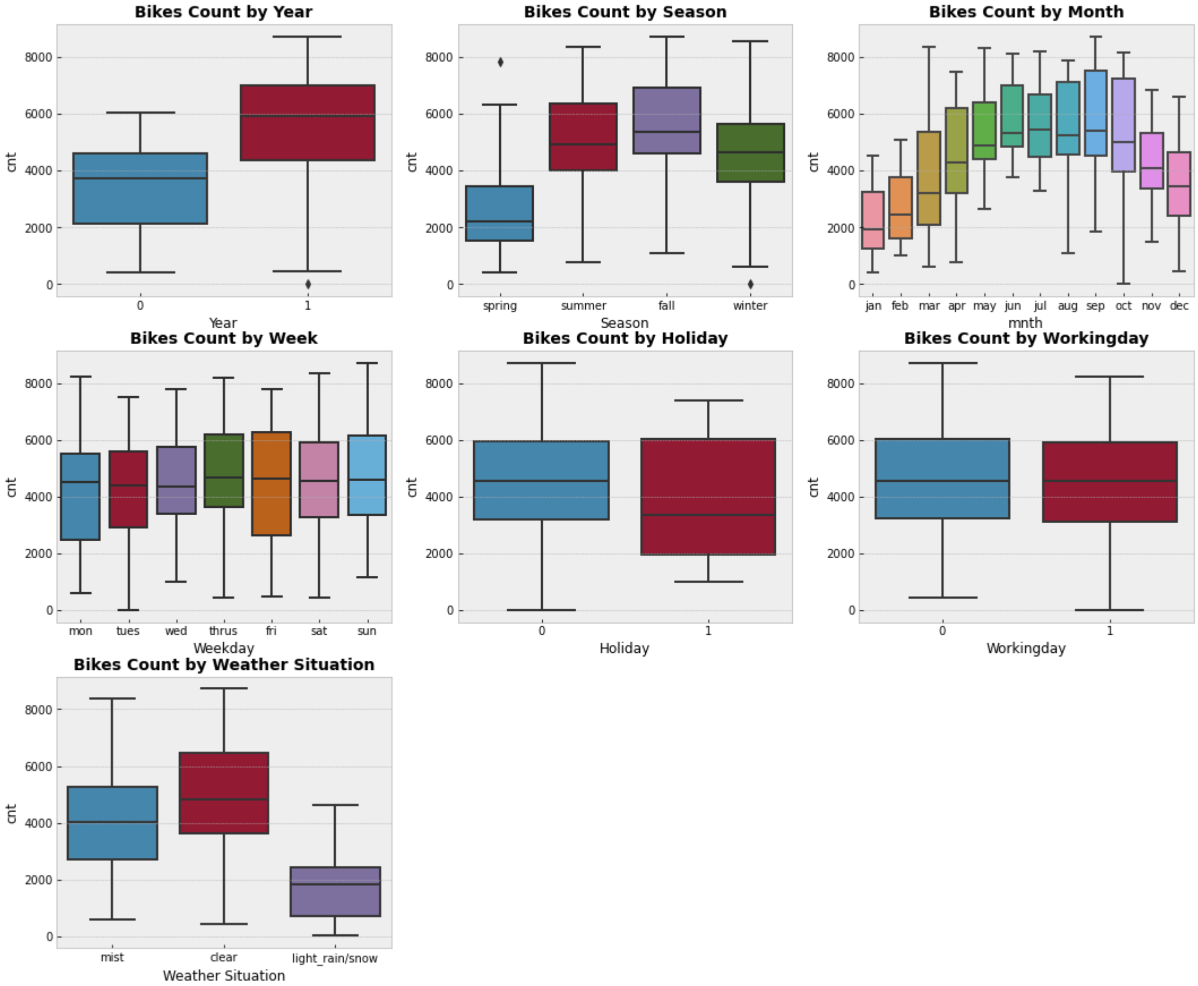
Subjective Questions

Ques 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans : From our analysis of the categorical variable from the data set. I am able to infer some of the following effects of categorical variables on the dependent variables :

- From the 'season' boxplot we can see that almost 5000 bookings are from the fall season compared to other seasons.
- From the 'yr' boxplot we can observe that the count of bike is increased in 2019.
- From the 'mnth' boxplot we can see that the months are following a trends and could be a good predictor varaible. The bookings in mid months are above 4000.
- From the 'holiday' boxplot most of the bike booking were happening when it is not a holiday. It means holiday cannot be a good predictor for the dependent variables.
- From the 'weekday' boxplot there are seems no trend in the weekday dataset, so we can leave for prediction.
- From the 'workingday' boxplot we can see that bike rental was on the higher end on days which were marked as non-working days. Also, the median count of bikes on working days equals the median count of bikes on working days.
- From the 'weathersit' boxplot we can see that the bike rental was on the higher end on days which were marked as clear, and also the median count of the bikes on clear days are greater as compare to any weather situation.

Please see the screenshot below :



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

Ans : A dummy variable is a numerical variable used in regression analysis to represent sub group of the sample.

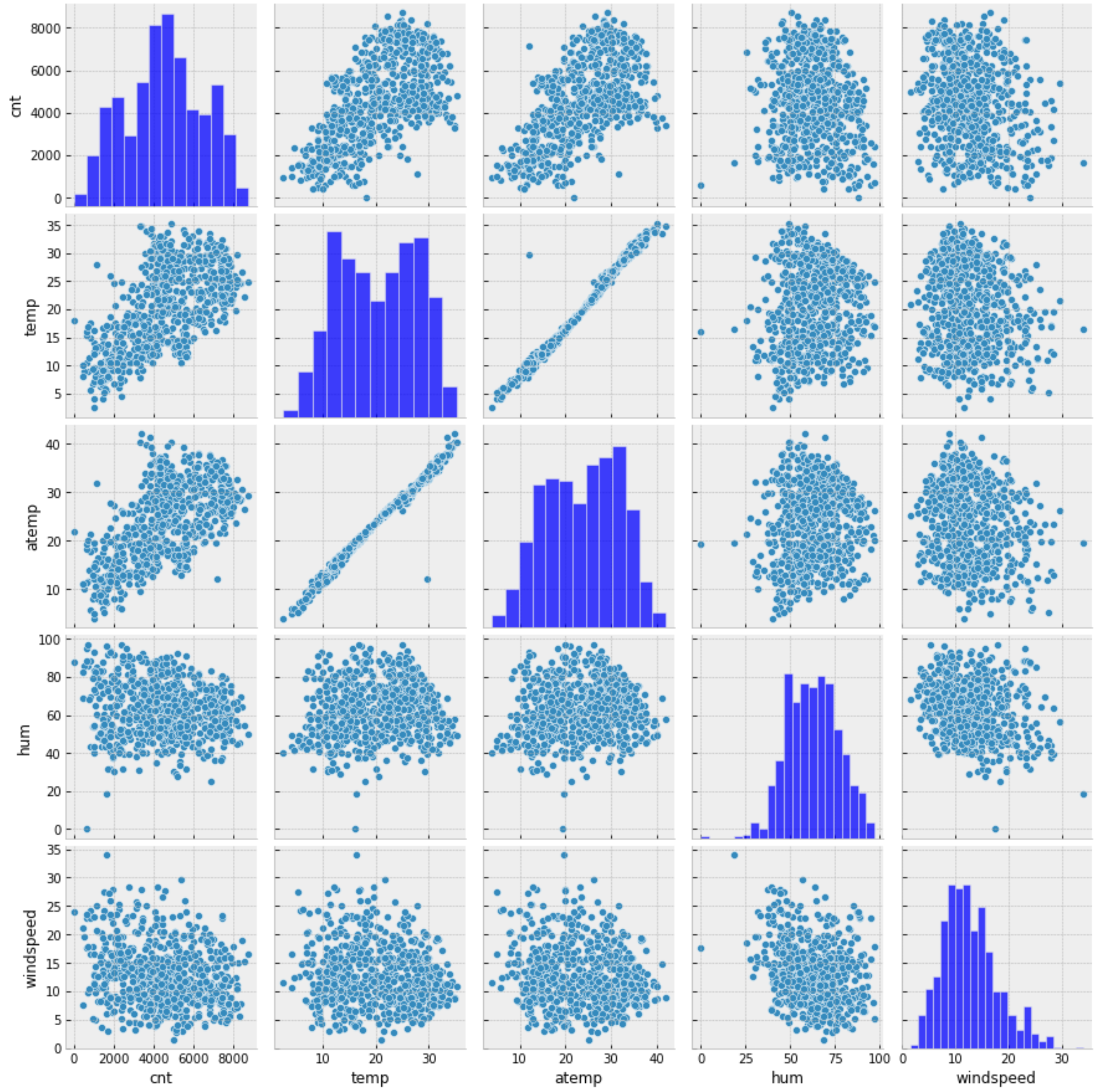
It is important to use drop_first = True, as it helps to reduce the extra column created during dummy variable creation. Hence it reduce the correlation created during dummy variables.

If we do not drop one of the dummy variables created from a categorical variables then it becomes redundant with data set as we will have constant variable(intercept) which will create multicollinearity issue.

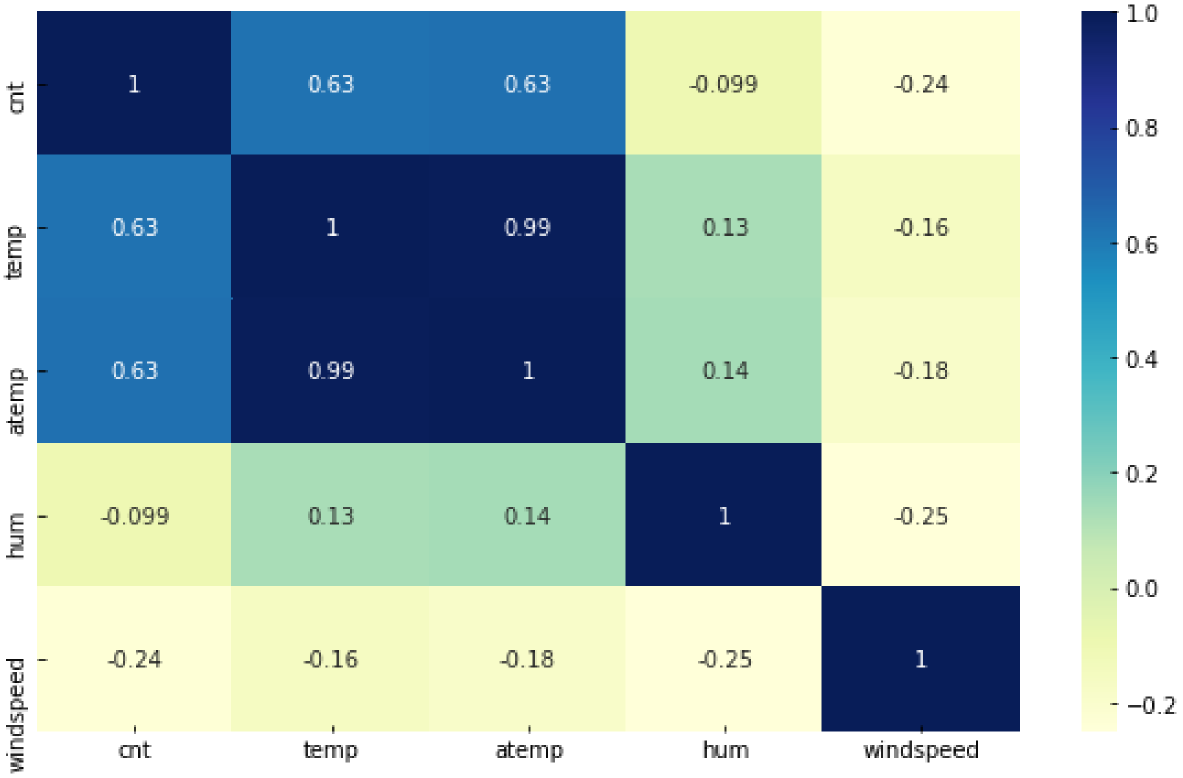
For example : Iterative models may have trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to multicollinearity between the dummy variables. To keep this under control, we lose one column.

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

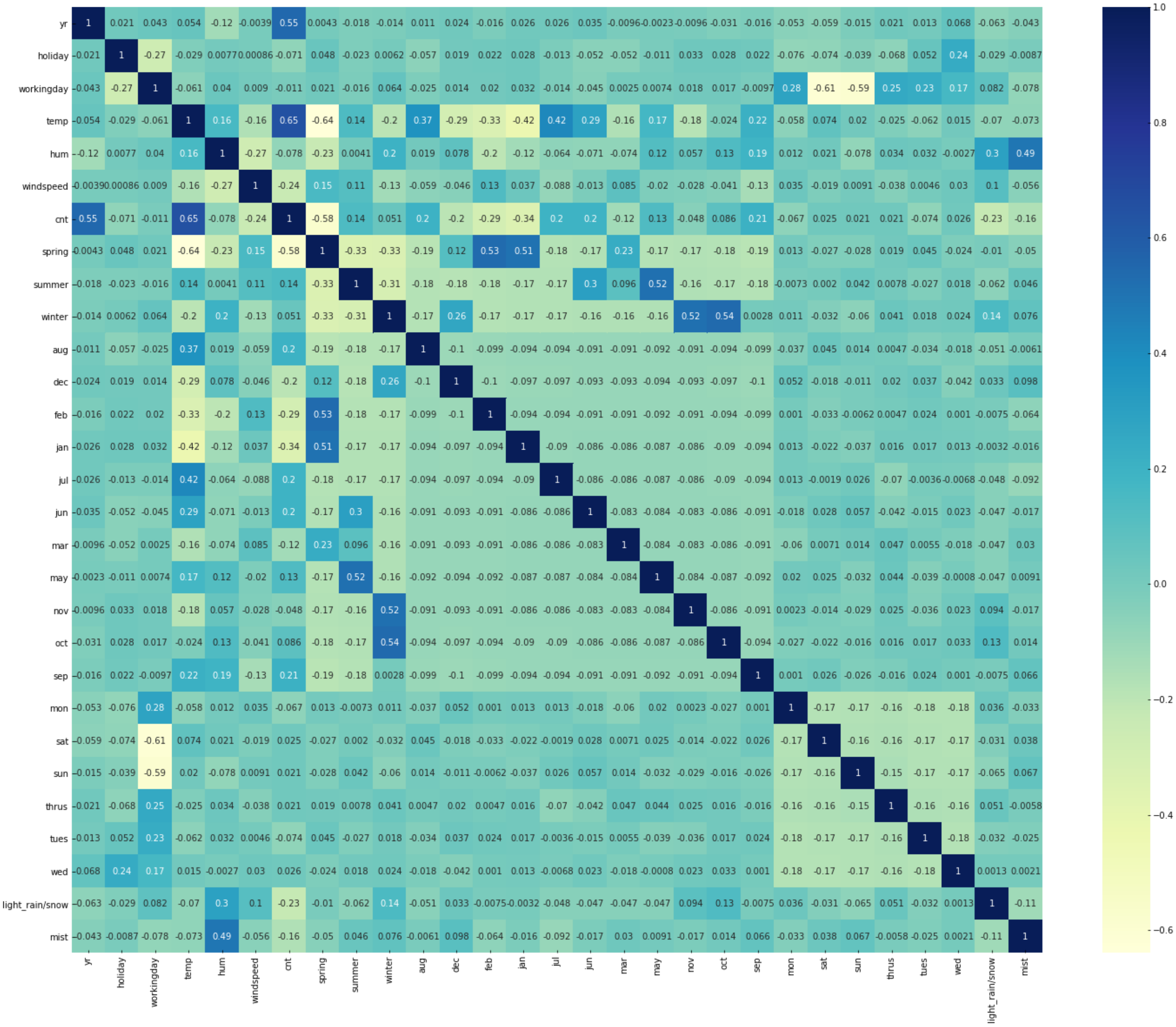
Ans : From the pair plot ‘temp’ has the highest correlation among the other numerical variables with ‘cnt’ as the target variable.



From the heatmap, it clearly shows which all variable.



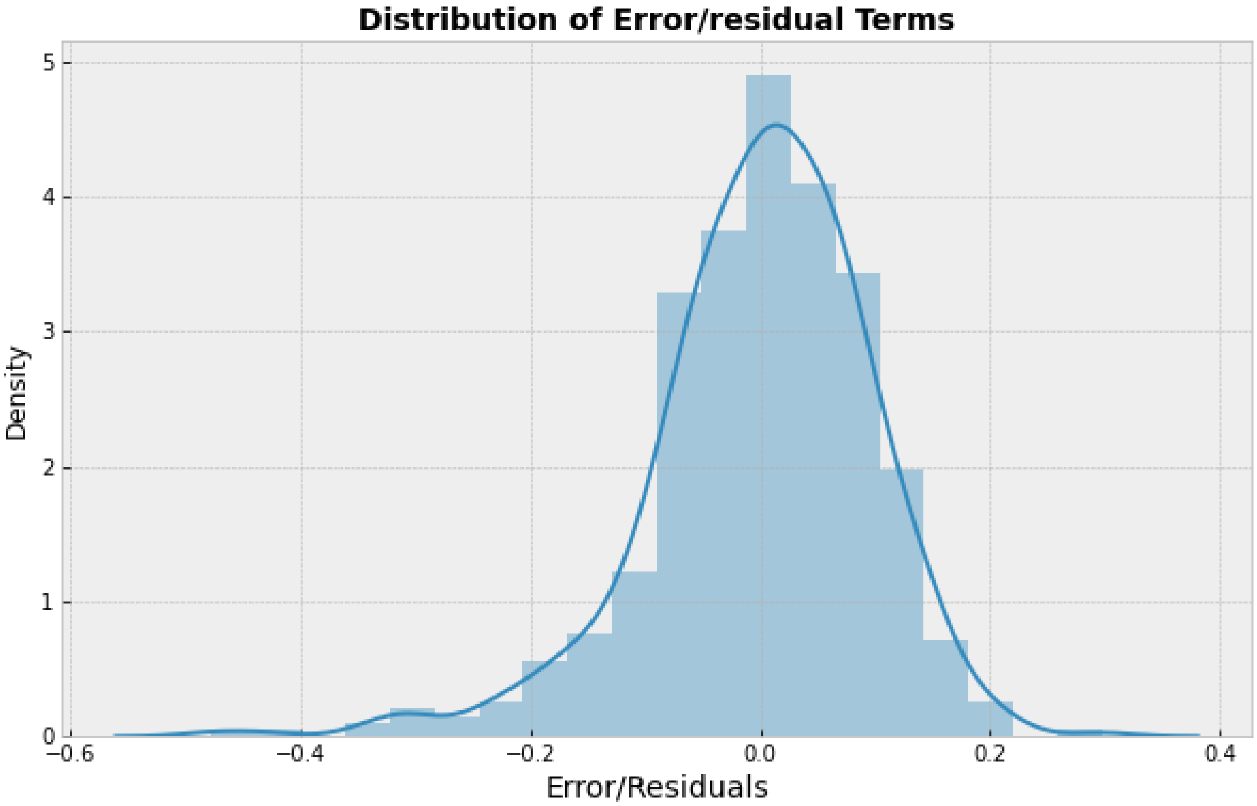
After creating Dummy variable. We observe target variable has highest correlation with temp. Please see the correlation (screenshot below):



	yr	holiday	workingday	temp	hum	windspeed	cnt	spring	summer	winter	aug	dec	feb	jan	jul	jun	mar	may	nov	oct	sep	mon	sat	sun	thrus	tues	wed	light_rain/snow	mist
yr	1	0.021	0.043	0.054	-0.12	-0.0039	0.55	0.0043	-0.018	-0.014	0.011	0.024	-0.016	0.026	0.026	0.035	-0.0096	-0.0023	-0.0096	-0.031	-0.016	-0.053	-0.059	-0.015	0.021	0.013	0.068	-0.063	-0.043
holiday	-0.021	1	-0.27	-0.029	0.0077	0.00086	-0.071	0.048	-0.023	0.0062	-0.057	0.019	0.022	0.028	-0.013	-0.052	-0.052	-0.011	0.033	0.028	0.022	-0.076	-0.074	-0.039	-0.068	0.052	0.24	-0.029	-0.0087
workingday	-0.043	-0.27	1	-0.061	0.04	0.009	-0.011	0.021	-0.016	0.064	-0.025	0.014	0.02	0.032	-0.014	-0.045	0.0025	0.0074	0.018	0.017	-0.0097	0.28	-0.61	-0.59	0.25	0.23	0.17	0.082	-0.078
temp	-0.054	-0.029	-0.061	1	0.16	-0.16	0.65	-0.64	0.14	-0.2	0.37	-0.29	-0.33	-0.42	0.42	0.29	-0.16	0.17	-0.18	-0.024	0.22	-0.058	0.074	0.02	-0.025	-0.062	0.015	-0.07	-0.073
hum	-0.12	0.0077	0.04	0.16	1	-0.27	-0.078	-0.23	0.0041	0.2	0.019	0.078	-0.2	-0.12	-0.064	-0.071	-0.074	0.12	0.057	0.13	0.19	0.012	0.021	-0.078	0.034	0.032	-0.0027	0.3	0.49
windspeed	-0.0039	0.00086	0.009	-0.16	-0.27	1	-0.24	0.15	0.11	-0.13	-0.059	-0.046	0.13	0.037	-0.088	-0.013	0.085	-0.02	-0.028	-0.041	-0.13	0.035	-0.019	0.0091	-0.038	0.0046	0.03	0.1	-0.056
cnt	0.55	-0.071	-0.011	0.65	-0.078	-0.24	1	-0.58	0.14	0.051	0.2	-0.2	-0.29	-0.34	0.2	0.2	-0.12	0.13	-0.048	0.086	0.21	-0.067	0.025	0.021	0.021	-0.074	0.026	-0.23	-0.16
spring	0.0043	0.048	0.021	-0.64	-0.23	0.15	-0.58	1	-0.33	-0.33	-0.19	0.12	0.53	0.51	-0.18	-0.17	0.23	-0.17	-0.17	-0.18	-0.19	0.013	-0.027	-0.028	0.019	0.045	-0.024	-0.01	-0.05
summer	-0.018	-0.023	-0.016	0.14	0.0041	0.11	0.14	-0.33	1	-0.31	-0.18	-0.18	-0.18	-0.17	-0.17	0.3	0.096	0.52	-0.16	-0.17	-0.18	-0.0073	0.002	0.042	0.0078	-0.027	0.018	-0.062	0.046
winter	-0.014	0.0062	0.064	-0.2	0.2	-0.13	0.051	-0.33	-0.31	1	-0.17	0.26	-0.17	-0.17	-0.17	-0.16	-0.16	-0.16	0.52	0.54	0.0028	0.011	-0.032	-0.06	0.041	0.018	0.024	0.14	0.076
aug	-0.011	-0.057	-0.025	0.37	0.019	-0.059	0.2	-0.19	-0.18	-0.17	1	-0.1	-0.099	-0.094	-0.094	-0.091	-0.091	-0.092	-0.091	-0.094	-0.099	-0.037	0.045	0.014	0.0047	-0.034	-0.018	-0.051	-0.0061
dec	-0.024	0.019	0.014	-0.29	0.078	-0.046	-0.2	0.12	-0.18	0.26	-0.1	1	-0.1	-0.097	-0.097	-0.093	-0.093	-0.094	-0.093	-0.097	-0.1	0.052	-0.018	-0.011	0.02	0.037	-0.042	0.033	0.098
feb	-0.016	0.022	0.02	-0.33	-0.2	0.13	-0.29	0.53	-0.18	-0.17	-0.099	-0.1	1	-0.094	-0.094	-0.091	-0.091	-0.092	-0.091	-0.094	-0.099	0.001	-0.033	-0.0062	0.0047	0.024	0.001	-0.0075	-0.064
jan	-0.026	0.028	0.032	-0.42	-0.12	0.037	-0.34	0.51	-0.17	-0.17	-0.094	-0.097	-0.094	1	-0.09	-0.086	-0.086	-0.087	-0.086	-0.09	-0.094	0.013	-0.022	-0.037	0.016	0.017	0.013	-0.0032	-0.016
jul	-0.026	-0.013	-0.014	0.42	-0.064	-0.088	0.2	-0.18	-0.17	-0.17	-0.094	-0.097	-0.094	-0.09	1	-0.086	-0.086	-0.087	-0.086	-0.09	-0.094	0.013	-0.0019	0.026	-0.07	-0.0036	-0.0068	-0.048	-0.092
jun	-0.035	-0.052	-0.045	0.29	-0.071	-0.013	0.2	-0.17	0.3	-0.16	-0.091	-0.093	-0.091	-0.086	-0.086	1	-0.083	-0.084	-0.083	-0.086	-0.091	-0.018	0.028	0.057	-0.042	-0.015	0.023	-0.047	-0.017
mar	-0.0096	-0.052	0.0025	-0.16	-0.074	0.085	-0.12	0.23	0.096	-0.16	-0.091	-0.093	-0.091	-0.086	-0.086	-0.083	1	-0.084	-0.083	-0.086	-0.091	-0.06	0.0071	0.014	0.047	0.0055	-0.018	-0.047	0.03
may	-0.0023	-0.011	0.0074	0.17	0.12	-0.02	0.13	-0.17	0.52	-0.16	-0.092	-0.094	-0.092	-0.087	-0.087	-0.084	-0.084	1	-0.084	-0.087	-0.092	0.02	0.025	-0.032	0.044	-0.039	-0.0008	-0.047	0.0091
nov	-0.0096	0.033	0.018	-0.18	0.057	-0.028	-0.048	-0.17	-0.16	0.52	-0.091	-0.093	-0.091	-0.086	-0.086	-0.083	-0.083	-0.084	1	-0.086	-0.091	0.0023	-0.014	-0.029	0.025	-0.036	0.023	0.094	-0.017
oct	-0.031	0.028	0.017	-0.024	0.13	-0.041	0.086	-0.18	-0.17	0.54	-0.094	-0.097	-0.094	-0.09	-0.09	-0.086	-0.086	-0.087	-0.086	1	-0.094	-0.027	-0.022	-0.016	0.016	0.017	0.033	0.13	0.014
sep	-0.016	0.022	-0.0097	0.22	0.19	-0.13	0.21	-0.19	-0.18	0.0028	-0.099	-0.1	-0.099	-0.094	-0.094	-0.091	-0.091	-0.092	-0.091	-0.094	1	0.001	0.026	-0.026	-0.016	0.024	0.001	-0.0075	0.066
mon	-0.053	-0.076	0.28	-0.058	0.012	0.035	-0.067	0.013	-0.0073	0.011	-0.037	0.052	0.001	0.013	0.013	-0.018	-0.06	0.02	0.0023	-0.027	0.001	1	-0.17	-0.17	-0.16	-0.18	-0.18	0.036	-0.033
sat	-0.059	-0.074	-0.61	0.074	0.021	-0.019	0.025	-0.027	0.002	-0.032	0.045	-0.018	-0.033	-0.022	-0.0019	0.028	0.0071	0.025	-0.014	-0.022	0.026	-0.17	1	-0.16	-0.16	-0.17	-0.17	-0.031	0.038
sun	-0.015	-0.039	-0.59	0.02	-0.078	0.0091	0.021	-0.028	0.042	-0.06	0.014	-0.011	-0.0062	-0.037	0.026	0.057	0.014	-0.032	-0.029	-0.016	-0.026	-0.17	-0.16	1	-0.15	-0.17	-0.17	-0.065	0.067
thrus	-0.021	-0.068	0.25	-0.025	0.034	-0.038	0.021	0.019	0.0078	0.041	0.0047	0.02	0.0047	0.016	-0.07	-0.042	0.047	0.044	0.025	0.016	-0.016	-0.16	-0.16	-0.15	1	-0.16	-0.16	0.051	-0.0058
tues	-0.013	0.052	0.23	-0.062	0.032	0.0046	-0.074	0.045	-0.027	0.018	-0.034	0.037	0.024	0.017	-0.0036	-0.015	0.0055	-0.039	-0.036	0.017	0.024	-0.18	-0.17	-0.17	-0.16	1	-0.18	-0.032	-0.025
wed	-0.068	0.24	0.17	0.015	-0.0027	0.03	0.026	-0.024	0.018	0.024	-0.018	-0.042	0.001	0.013	-0.0068	0.023	-0.018	-0.0008	0.023	0.033	0.001	-0.18	-0.17	-0.17	-0.16	-0.18	1	0.0013	0.0021
light_rain/snow	-0.063	-0.029	0.082	-0.07	0.3	0.1	-0.23	-0.01	-0.062	0.14	-0.051	0.033	-0.0075	-0.0032	-0.048	-0.047	-0.047	-0.047	0.094	0.13	-0.0075	0.036	-0.031	-0.065	0.051	-0.032	0.0013	1	-0.11
mist	-0.043	-0.0087	-0.078	-0.073	0.49	-0.056	-0.16	-0.05	0.046	0.076	-0.0061	0.098	-0.064	-0.016	-0.092	-0.017	0.03	0.0091	-0.017	0.014	0.066	-0.033	0.038	0.067	-0.0058	-0.025	0.0021	-0.11	1

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

Ans : The assumption that we make after we building the linear regression model on the training data set is that error terms are distributed normally. In support of that we have do the residual analysis. Residual represents the error of the different between actual y values and predicted y value by the model.



From the above diagram as we could see that the residual are normally distributed and maximum of the error terms resolving around zero. Hence our assumption of linear regression is valid. Also ensured the overfitting by looking the R-square value and Adjusted R-square value. We also validate this assumption about residuals by plotting a displot of residual and see if residuals are following normal distribution or not.

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

Ans : From the final model the top 3 variables that needed for the prediction which are influence the counts are :

- **Temperature (temp)** : The demand for bikes rises with the increase in temperature. For every unit increase in temperature, the bike hire count increases by 0.3827 units, when all other variables are kept constant.
- **Year(yr)** : As the year variable increases, there is an increasing demand of bikes. For every increase of one year, the bike hire count is expected to increase by 0.2315 units, keeping all other variables as constant.
- **Winter** : For every unit increase in the winter variable,the count is expected to increase by 0.0744 units, when calculated with respect to Fall season as reference , keeping all other variables as constant.

General Subjective Questions

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

Ans : Linear regression is a supervised machine learning algorithm where the predicted output is continuous and has constant slope. It’s used to predict values within a continuous range (e.g. sales, price) rather than trying to classify them into categories.

Linear regression is one of the basic forms of machine leaning which we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linear correlated.

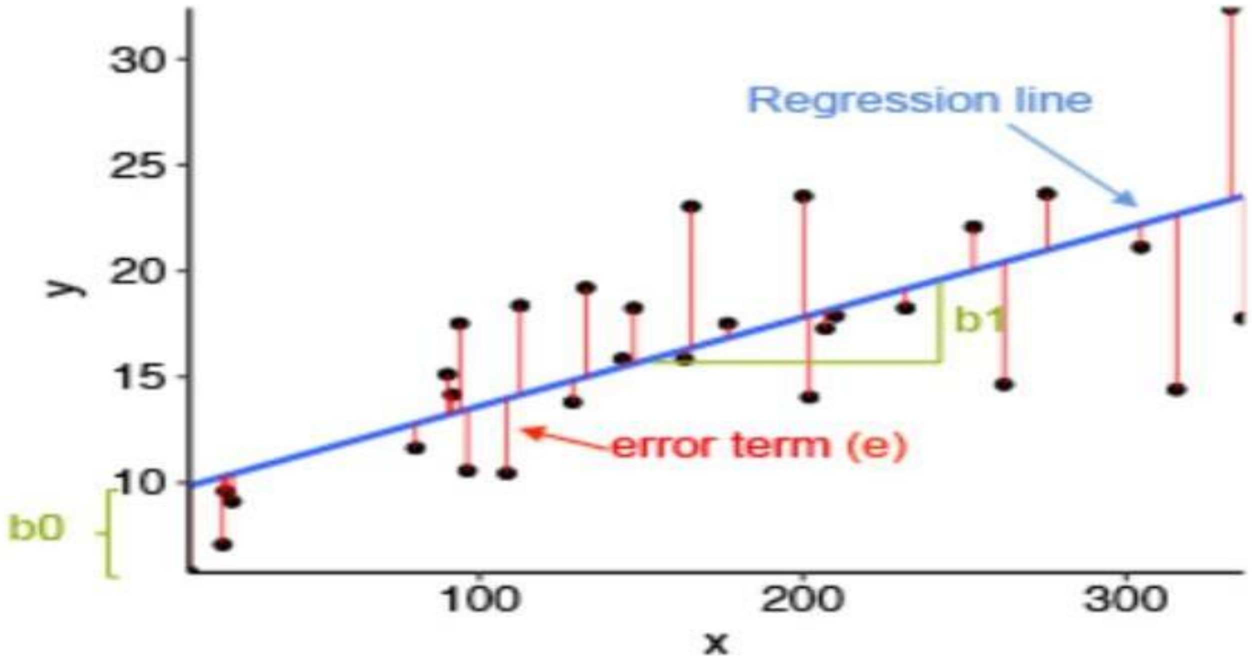
Mathematically, we can write a simple linear regression equation as follow :

$$\mathbf{Y = m*x + c}$$

Here,

- y : is the predicted variable (dependent variable),
- m is slope of the line,
- x is independent variable,
- c is intercept(constant).

It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.



Here x and y are the two variables on the regression line.

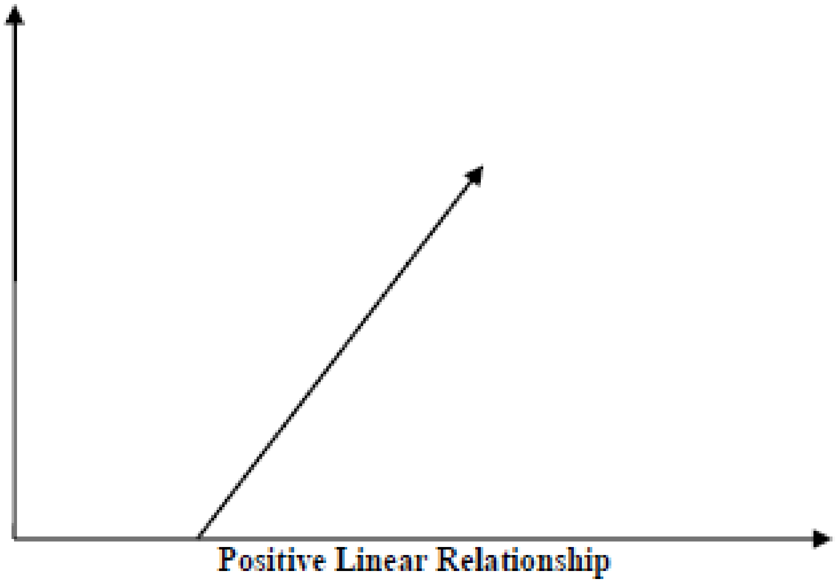
b1 – slope of the line, and **b0** – y-intercept of the line

x – independent variable, and **y** – dependent variable from the data set.

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

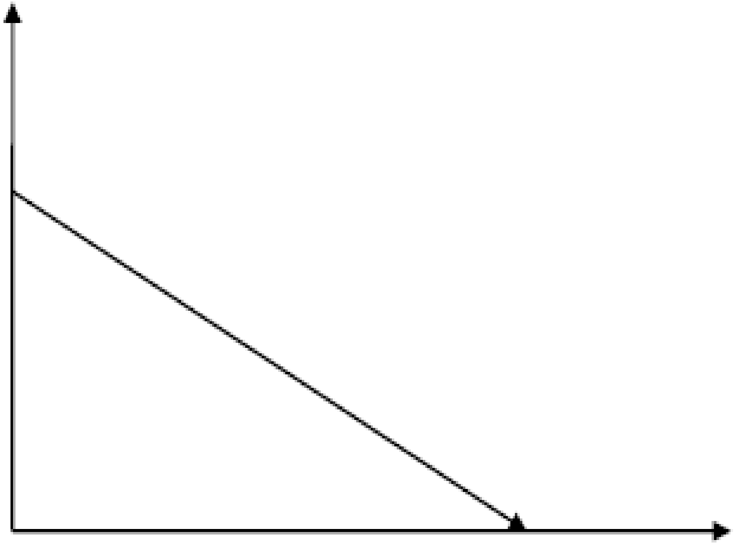
Positive Linear Relationship:

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



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 –

- **Simple Linear Regression** –
  It explains the relationship between a dependent variable and only one independent variable using a straight line.
  Formula: $Y=\beta 0+\beta 1X1 +\epsilon$

- **Multiple Linear Regression**-
  It shows the relationship between one dependent variable and several independent variables.
  Formula: $Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$

Assumptions :

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

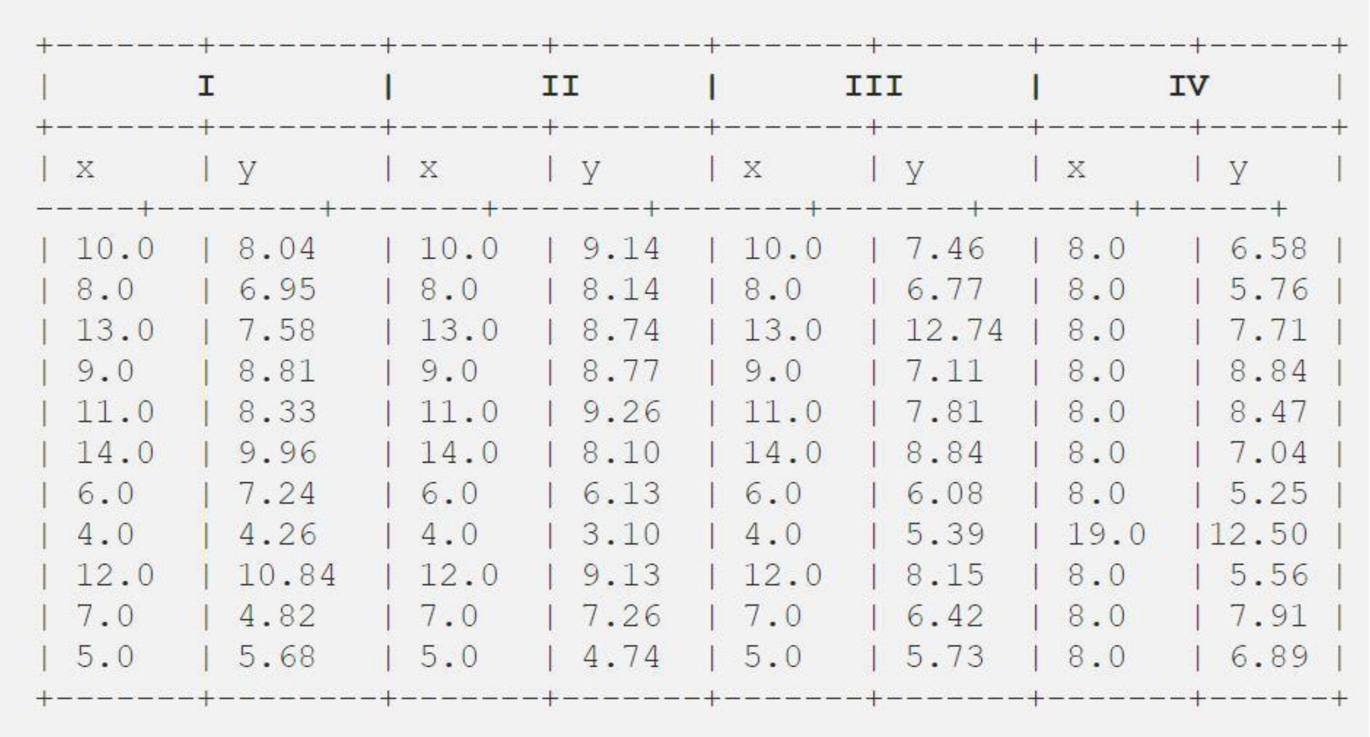
- **The assumption about the form of the model:**
  It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption.

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

Ans : 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.

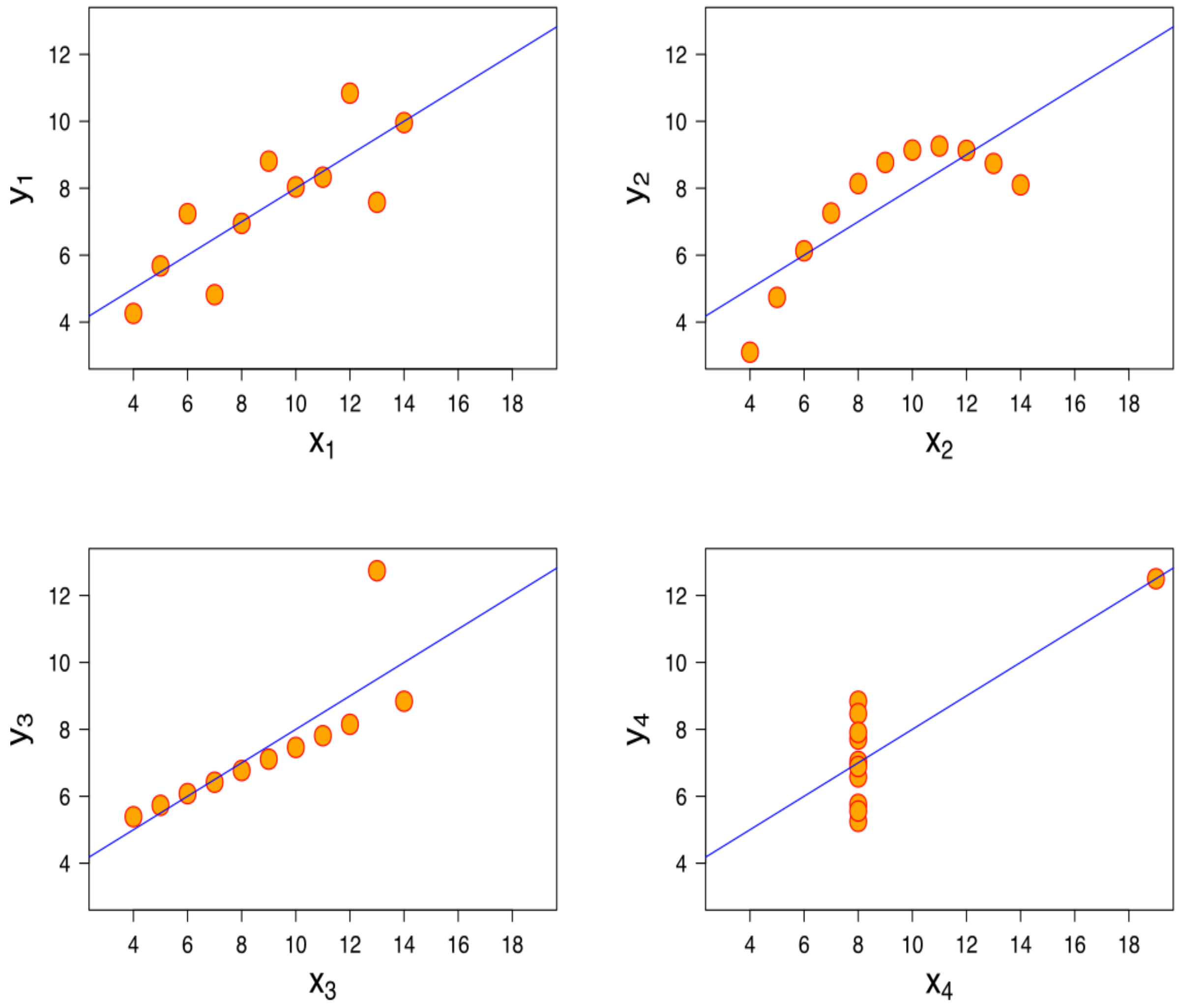
Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation and correlation between x and y. 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:



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

Ques 3 : What is Pearson'R ? (3 marks)

Ans : In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between **−1** and **1**.

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)

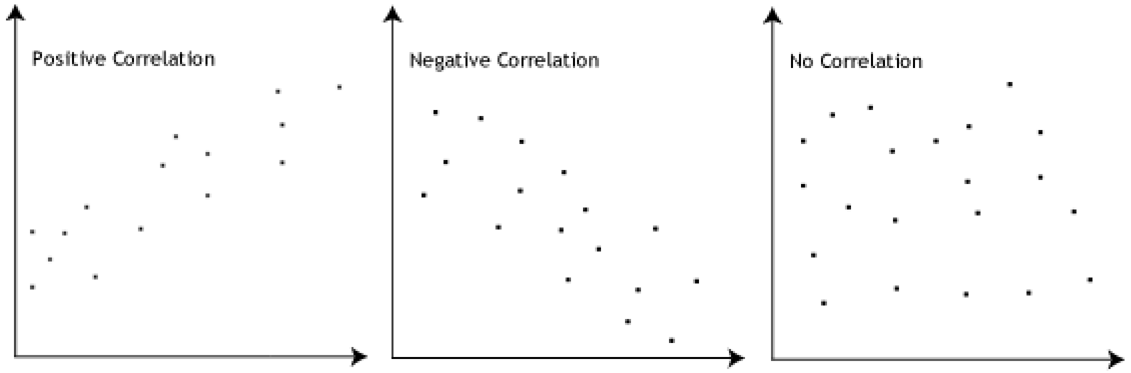
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

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



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means for every positive increase in one variable, there is a negative increase of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

- Zero means that for every increase, there isn’t a positive or negative increase. The two just aren’t related.

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

Here,

R = Correlation coefficient

xi=values of the x-variable in a sample

$\bar{x}$=mean of the values of the x-variable

$\bar{y}$=values of the y-variable in a sample

yi=mean of the values of the y-variable

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

Ans : **Scaling:** Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

Why is scaling performed:

1. Machine learning algorithm just sees number, if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

2. To ensure that the gradient descent moves smoothly towards the minima

and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Normalization scaling :

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.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- 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
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization scaling :

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:

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

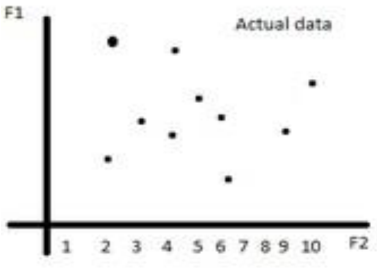
μ is the mean of the feature values σ and is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

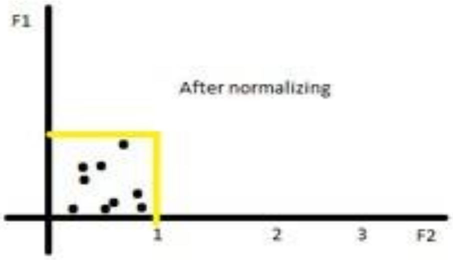
Difference Between Normalization and Standardization:

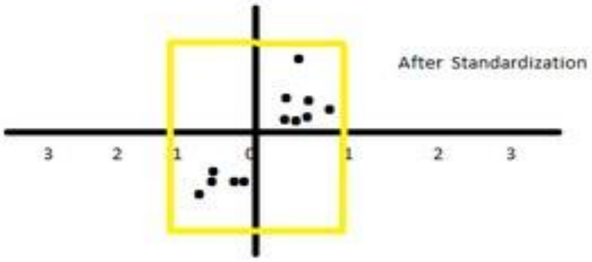
- Normalization is good to use when we know that the distribution of our data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in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

Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unit less. Refer to the below diagram.

Below diagram shows that hoe the data looks like after scaling:







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

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 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.

An infinite VIF value indicates that the dependent variable may be expressed exactly by a linear combination of other variables. VIF =1/(1-R2), where E2=1 then VIF = Infinity.
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:

- ➢ 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.

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

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

Ans :Q-Q plot is a probability plot, which is a graphical method for comparing two probability distribution by plotting their quantile against each other. First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

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 It's being compared to a set of data on the y-axis.

Advantages:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

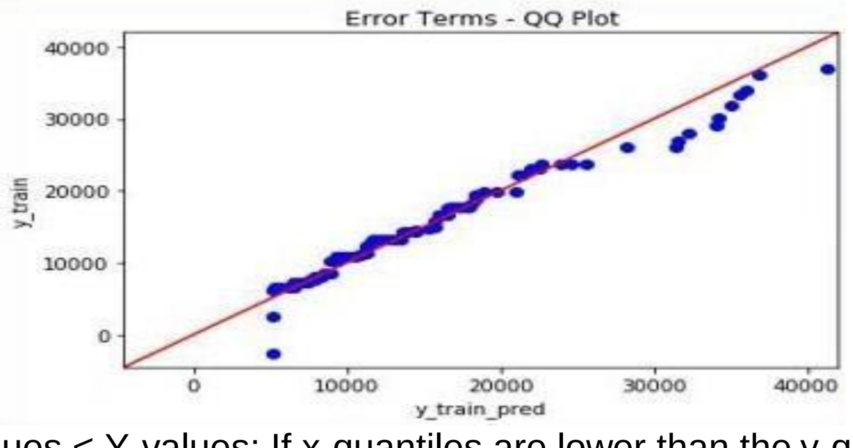
The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?

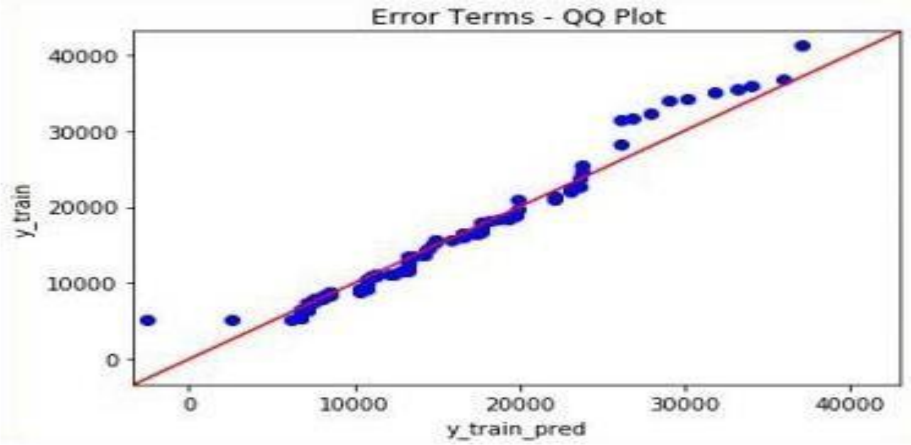
Below are the possible regressions for two data sets:

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

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



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



4. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.