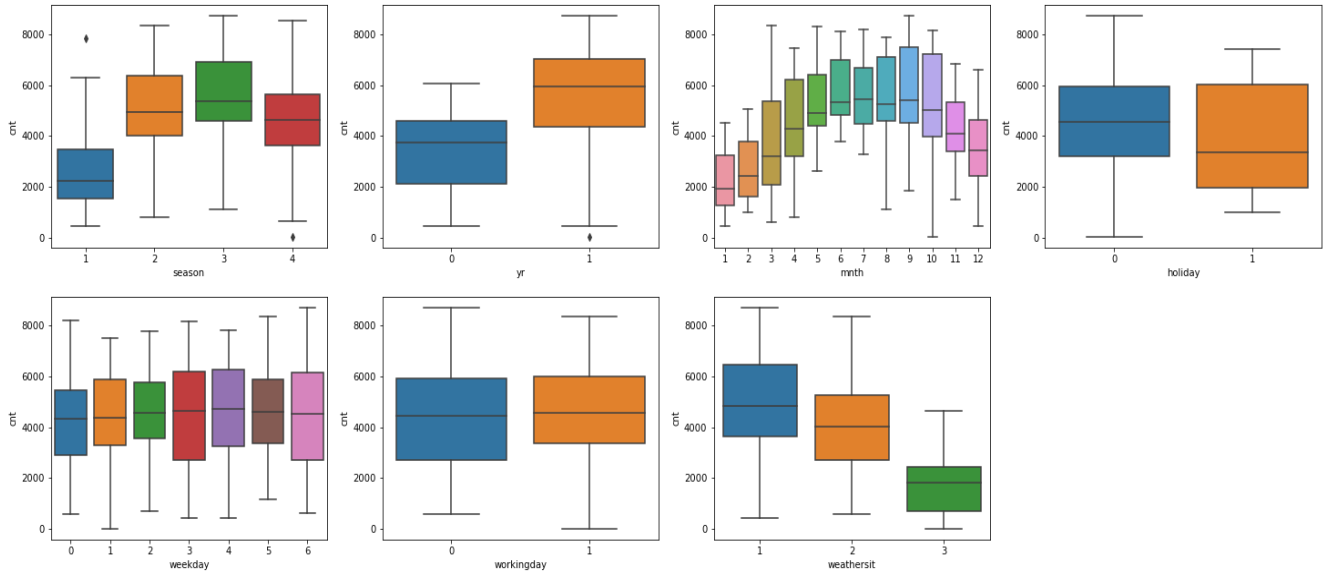
Assignment-based Subjective Questions

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



The following information can be inferred from the categorical plot,

1. In first plot, the maximum number of bookings around 5000 in season_3 which is “Fall”.
2. In second plot, the number of bookings is increased in year_1(2019) compared to previous year (2018).
3. In third plot, the number of bookings is gradually increased from Month_1 to month_6 which is “Jan to June” and it decrease gradually. The maximum number of bookings is on month-6 which is “June”
4. In fourth plot, when there is a holiday, the demand is decreased compared to working day.
5. In plot five and six, there is no significant insights that we can make.
6. In seventh plot, the count is increasing when the weather is Clear, Few clouds, partly cloudy, partly cloudy. There are no bookings when the weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

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

In order to eliminate the correlation / redundant while creating dummy variables, we need to use drop_first. In Linear regression, we need to convert the categorical variables into binary (0 and 1) representation. If we have N categorical variables then we need to create N-1 dummy variables.

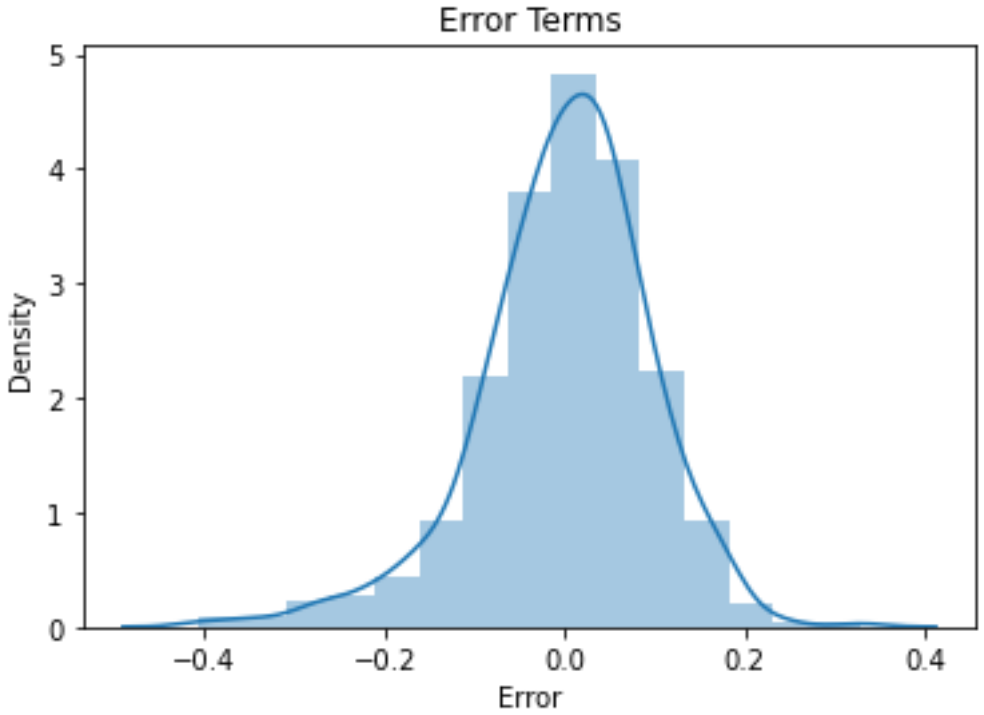
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

By looking at the pair plot “Temp” and “atemp” are highly correlated with the target variable “cnt”.

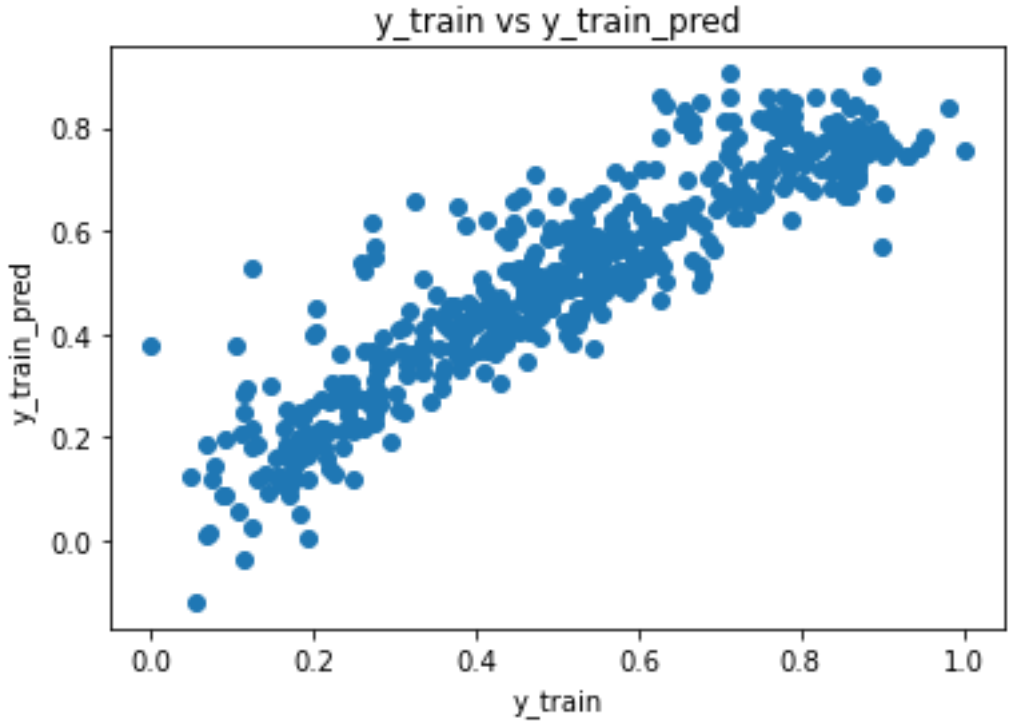
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The assumptions of Linear Regression are,

1. By using pair plot, we can find the linear relationship between the independent and target variables.
2. Residual Analysis – Error terms are normally distributed with the mean is equal to Zero.



3. By using scatter plot, we can find the error terms have constant variance (Homoscedasticity)



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

According to my final model below 3 features contributing significantly towards the demand of the bike sharing,

1. Temp – Coefficients of temp is 0.5288. It means that if temperature increases the demand also increase.
2. Year – Coefficient of year is 0.2301. It means that the unit increase in Year variable, the demand of bike also increases.
3. Windspeed – Coefficient of windspeed is -0.1189. It means that the unit increase in windspeed variable, the demand of bike is decrease.

General Subjective Questions

1.Explain the linear regression algorithm in detail

Linear Regression algorithm is based on supervised machine learning algorithm. It is the part of Regression analysis. Regression analysis is the technique to find out the relationship between the input and target variables. The outcome of the regression analysis is numerical / continuous. Eg: Predicting scores, salary.

Two types of regression,

1. Simple Linear Regression: It has only one independent variables with respect to dependent variables.
2. Multiple Linear Regression: It has more than one independent variables with respect to dependent variables.

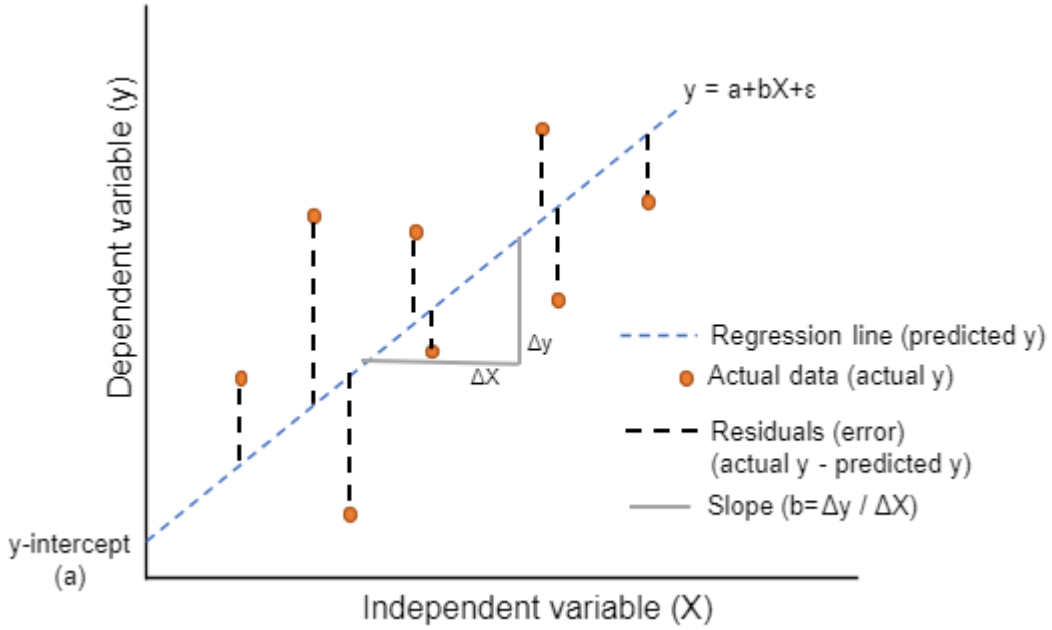
Basic assumptions of the linear regression model,

1. Linear relationship between the target and independent variables.
2. Error terms are normally distributed with mean equal to Zero.
3. Error terms are independent to each other
4. Error terms have constant variance.

The mathematical expression for linear regression is

$$y = B0 + B1*x + e$$

where B0 = Intercept, B1 = Slope, e= Error Term



In above picture, the blue dotted line shows the best-fit line.

After we built the linear regression, we need to calculate the R squared value,

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R^2_{adj} = 1 - ((1 - R^2) \frac{N-1}{N-M-1}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

Were,

RSS – Residual Sum of Square – Difference between the actual Y and Predicted Y variable

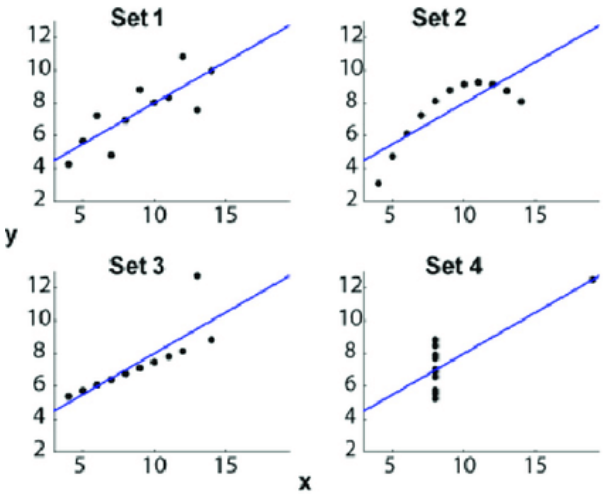
TSS – Total Sum of Square – Difference between the actual Y and average Y values.

N- Number of Rows, M – Number of Columns

Multi-collinearity issues will come under multi-Linear regression algorithm. To avoid multi-collinearity, we need to check the p-value and VIF (Variable Inflation Factor). When p-value is >0.05, we need to remove the features and if VIF > 5 on the linear regression, we need to remove the feature as well.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets which are nearly identical in simple statistical properties but it appears differently on scatter plots.



Mean of Y	**7.50 in all 4 XY plots**
Sample variance of Y	**4.122 or 4.127 in all 4 XY plots**
Correlation (r)	**0.816 in all 4 XY plots**
Linear regression	**y = 3.00 + (0.500 x) in all 4 XY plots**

Data sets for the 4 XY plots

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

The above picture shows four dataset and corresponding mean, variance is same for all 4 datasets. But when we visualize the dataset in scatter plot, all 4 plots are completely different and that is not interpretable by any algorithm which is fooled by these peculiarities.

Set 1 – Shows the perfect relationship between x and y is Linear

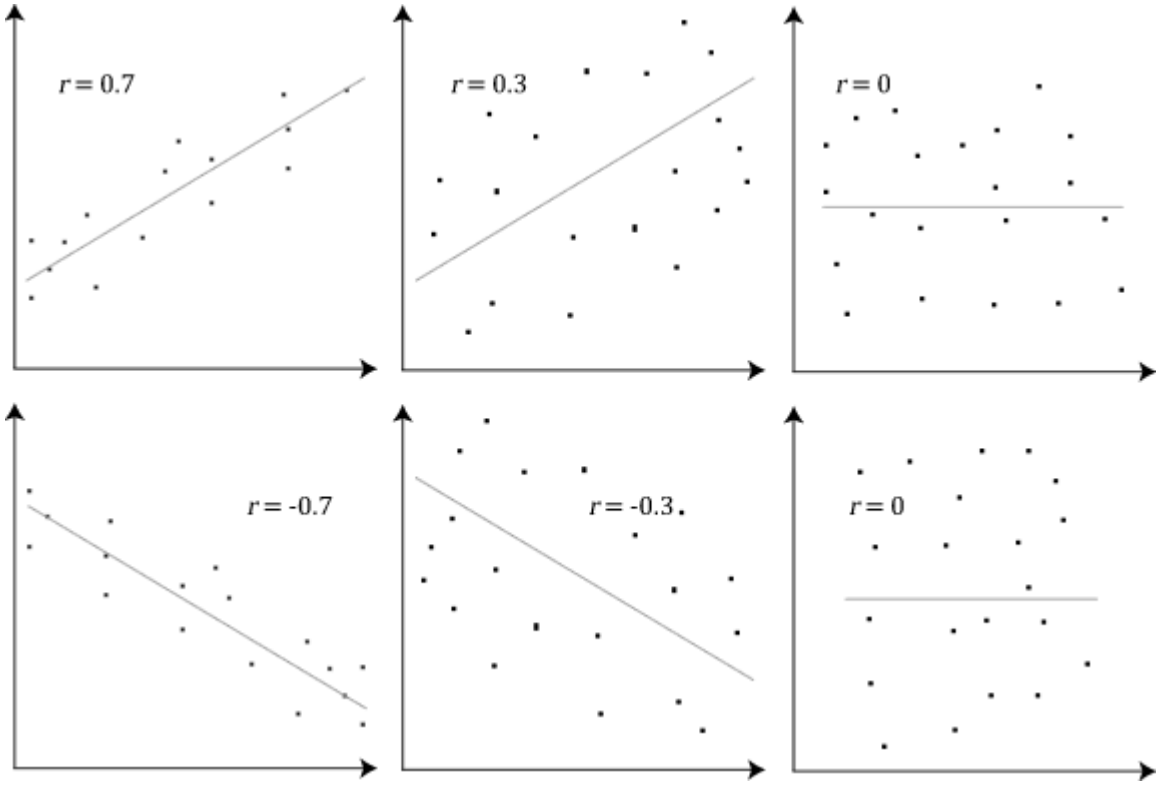
Set 2 – shows the relationship between x and y is non-linear

Set 3 – shows the relationship between x and y is Linear but except one dataset which is far away from the line which seem to be Outlier.

Set 4 – It has one high-leverage point which is enough to get the high correlation coefficient.

3. What is Pearson's R?

Pearson's R is known as Pearson's correlation coefficients (denoted as r). It is used to measure the strong relationship between the variables x and y. Different types of correlation coefficients are present, but Pearson's R is most popular and mostly used in Linear regression. The coefficients range from -1 to 1.



The above diagram shows the different types of correlation.

1. r=0.7 and r = 0.3 has Positive correlation between x and y. if r = 1 means, the data points are included in the best fit line. If 0>r<1 means, the data points have some variance around the line.
2. r= -0.7 and r = -0.3 has Negative correlation between x and y. if r = -1 means, the data points are included in the best fit line. If -1>r<0 means, the data points have some variance around the line.
3. r=0 means there is No correlation between the variables x and y.

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

It is the technique to use standardise / normalize the independent features present in the data to particular range. In Linear regression, we need to perform this technique during data pre-processing to handles the units or values.

In Machine learning algorithms like Linear, logistic regression, we use gradient descent to find the optimal solution. The difference in the range of features will cause different step size of each feature. In order to obtain the smooth gradient descent towards local minima, we need to update all the features has same scaling.

Example: If one of the feature units as km and value is 1 km and other variable unit as m and value is 1000m. We know mathematically both the values are same. But if we didn't do the scaling, we will get different step size for gradient descent.

Two types of feature scaling as

1. Min-Max (Normalisation) scaling
2. Standardisation scaling

In Normalisation scaling, it brings all the data are present in the range between 0 to 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

In Standardisation scaling, it brings all the data into standard normal distribution with mean zero and standard deviation as 1.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

VIF is a measure of multi-collinearity of a variable in the datasets. In regression model, some of the features are interrelated and it cause the model is redundant. In order to avoid these features, we use VIF to find out the correlation between other variables. If VIF is high means, the variables have strong correlation with other features in the datasets.

$$VIF = \frac{1}{1 - R^2}$$

If VIF is 1 means, predictor is not correlated with other variables. If the values of VIF as less than 5, which means that moderate correlation with other variables.

Sometimes VIF as Infinite which means that perfect correlation between two independent variables.

In this case R^2 value will be 1, so VIF as Infinite.

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

Q-Q plot (Quantile- Quantile plot) are plots of two quantiles against each other. It is graphical tool to asses if the set of data are from theoretical distribution such as normal or exponential or uniform distribution.

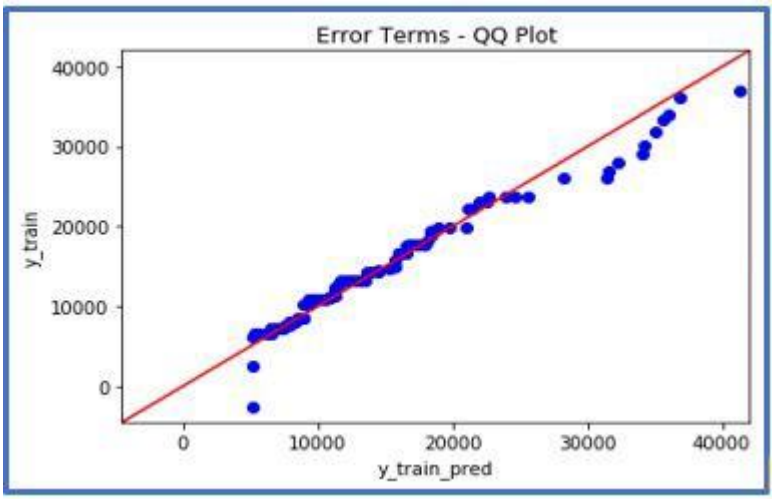
This plot helps in the linear regression when the training and testing dataset received separately, we can use Q-Q plot to confirm both the datasets are from population with common distribution.

Few advantages,

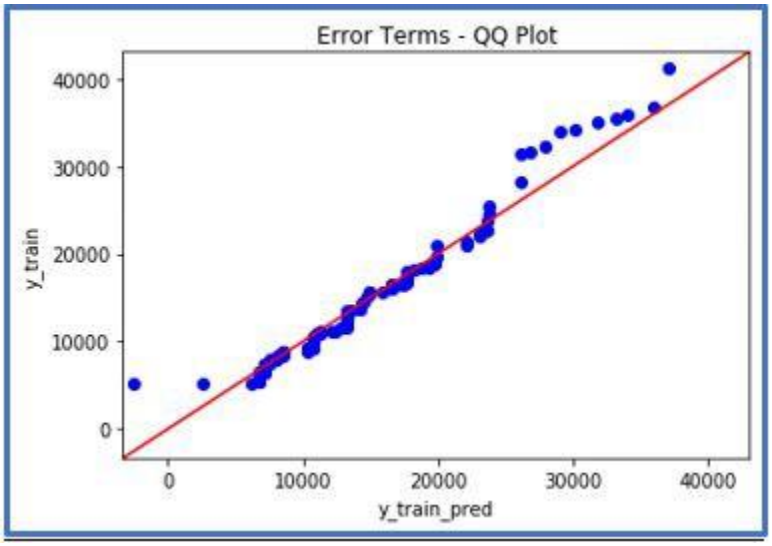
1. It can use same sample size
2. Many distributional aspects like shifts in scale, shifts in location and presence of outliers also detected in this plot

Below are the possible interpretations of the two datasets: -

1. Similar distribution – If all the points of quantiles lie on or close to the straight line at an angle of 45 degree from X-axis.
2. Y- values < X- values – If Y quantiles are less than X quantiles



3. X - values < Y- values – If X quantiles are less than Y quantiles



4. Different distribution - – If all the points of quantiles lie away from the straight line at an angle of 45 degree from X-axis.