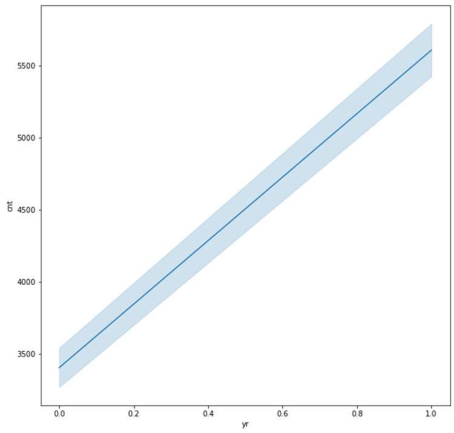
ASSIGNMENT BASED SUBJECTIVE QUESTIONS

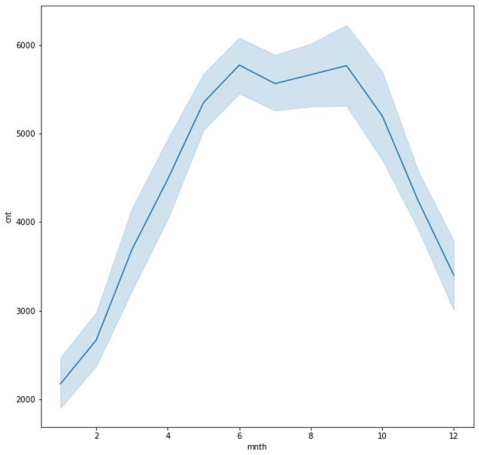
Q. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A. Following categorical variables are observed in our database which are in numerical form and their independent effect on the dependent variable ('cnt') is as below:-

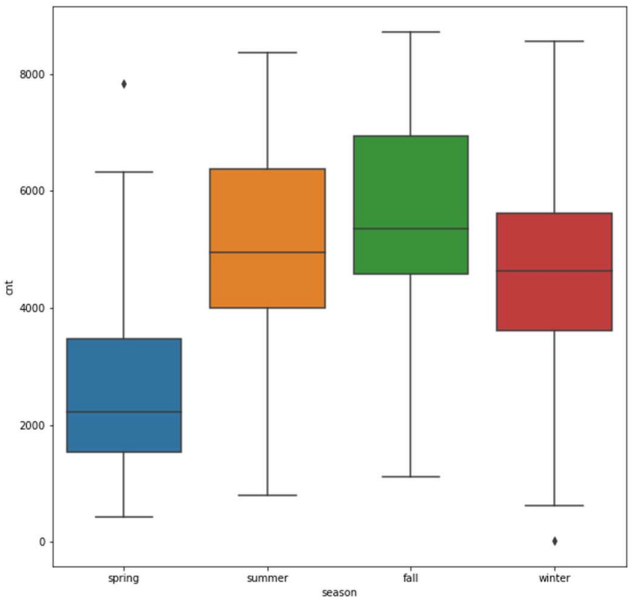
1. yr - cnt shows an overall increase YoY based on the 2 years data that we have. This can be seen from the below line plot.



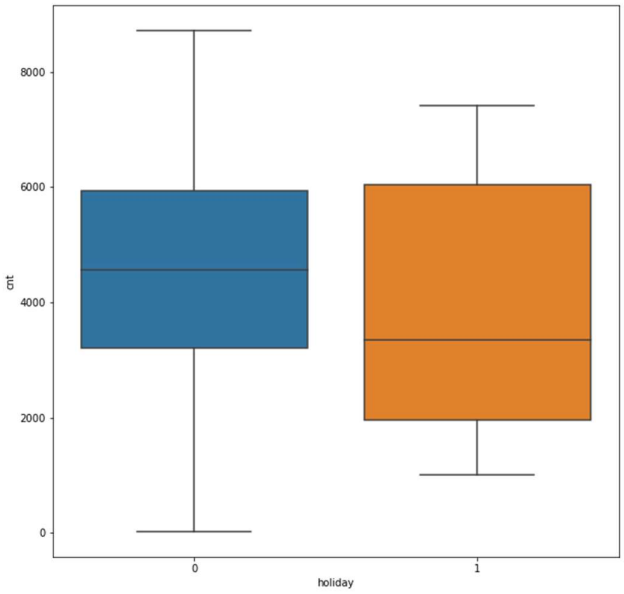
2. mnth – there is gradual increase in demand from month of January and there is peak demand between the months of May to October after which the demand is again seen to be decreasing. This can be seen from the below line plot.



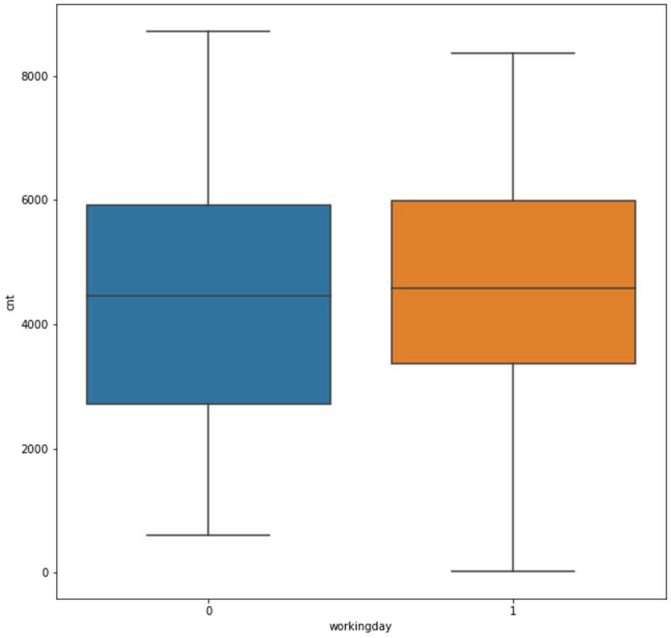
3. season – there is peak demand during the fall season followed by marginal decrease in overall demand during the summer season. Winter season again has less demand then summer season and during spring season there is least demand. This can be seen from the below box plots.



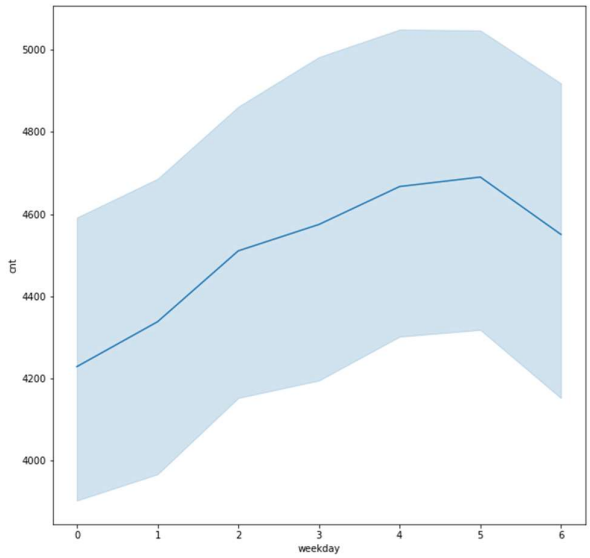
4. holiday – the peak demand remains same if there is holiday or no holiday. However in general there is tendency to have a low demand during holidays which can be inferred from the below box plot. (0 indicates no holiday and 1 indicates holiday)



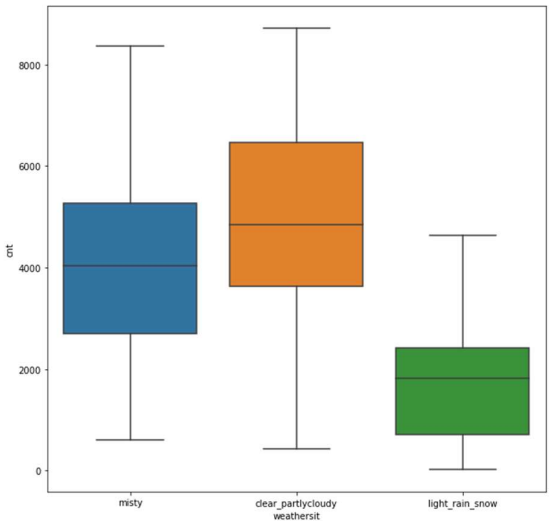
5. Workingday – there seems to be a marginal decrease in the demand during non-working days rather than working days. This can be seen from the below box plot.



6. Weekday – On comparing the columns weekday and workingday it is assumed that weekday 0 indicates a Sunday and weekday 6 indicates a Saturday. Assuming 5 days a week working, If we see the below line plot we can observe that there is decrease in ‘cnt’ during Sundays and Saturdays which somewhat matches our inferences of target variable w.r.t holidays. During working day i.e. from Monday to Friday, the demand increases from Monday and goes to peak on Friday



7. Weathersit – From the below box plot we can observe that people like to use this app during clear / partly cloudy days. There is decrease in demand during misty days and very low demand during mild rains / snowy days.



Q. Why is it important to use Drop_First = True during dummy variable creation

A. If there is a categorical variable with “k” distinct values which are of unordered nature, then these can be converted to dummy variable with column name as ‘value name’ and value as 1 if column is true / present and 0 if column is false / absent.

When such dummy variables are being created , for a categorical variable with ‘k’ distinct values, k=1 such variables should be created since one of the variable can be completely represented by these other dummy variables.

For e.g. in this case for the unordered categorical variable “season” had 4 distinct values “spring” , “fall”, “summer” and “winter”, we created 3 dummy variables namely “spring” , “fall” and “summer”

Spring	Fall	Summer	RESULTANT VALUE
1	0	0	Spring
0	1	0	Fall
0	0	1	Summer
0	0	0	Winter

As seen above the season winter can be represented when the 3 other dummy variables carry the value 0.

So while using the get_dummies() function the argument Drop_First drops the first value and keeps only the remaining values to create the dummy variables.

Hence it is important to use the Drop_First = True argument.

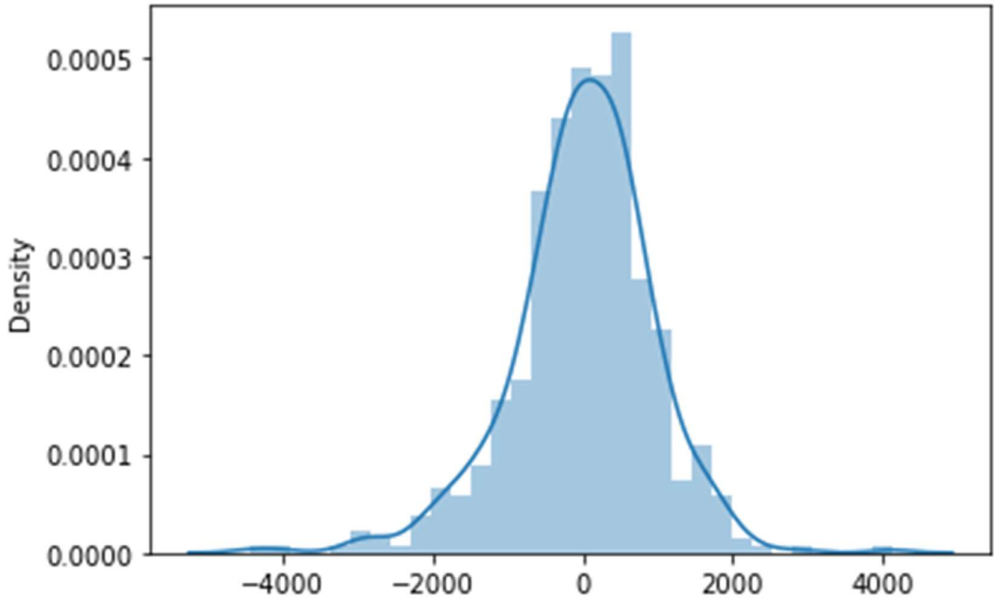
Q. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A. The variable ‘registered’ has the highest correlation with the target variable with the correlation value of 0.95

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

A.

1. **Distribution of errors is normal with mean approximated along zero:** This is checked by computing the residue values (y_train – y_test_predicted) and plotting a histogram of the residue values. This is found as below which validates this assumption



2. **Predictor variables are not strongly correlated with each other (No Multicolinearity) :** This is checked by checking the VIF of the final model which shows that the VIF of all the variables which are taking a part in our final model is less than 5 indicating very low multicolinearity.

	Features	VIF
4	workingday	4.530
1	mnth	4.094
10	reg_by_casual	3.743
3	weekday	3.258
5	spring	2.148
0	yr	1.984
6	summer	1.763
7	fall	1.670
8	misty	1.608
2	holiday	1.134
9	light_rain_snow	1.133

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

A. The below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

a. year (positive coefficient)

b. whether the season is spring (negative coefficient)

c. whether the weather is light rainy or light snowy (negative coefficient.

GENERAL SUBJECTIVE QUESTIONS

Q. Explain linear regression algorithm in detail

A. Regression, in general, is defined as relation between a target variable and the other input variables.

When the input variables are seen to follow a linear relationship with the output variable, then linear regression algorithm can be used.

Consider a graph between X(Independent variable) and Y(Dependant Variable) on which a straight line is drawn.

The straight line will follow the equation

$Y = mX + C$

M: Slope of the line

C: Intercept of the line

With this equation, at any point, value of Y can be predicted with value of X.

Now let us consider an example of database which has 2 columns

a. Work experience
b. Annual Salary

Let us consider that the annual salary increases linearly with the work experience.

However when we will take practical values of annual salaries w.r.t work experience they will not exactly follow a linear relationship but will have points that will be scattered along some undefined / unknown line which can define the linear relationship between the work experience and the salary.

Computation of this undefined line from the data that we have can be described as linear regression.

With linear regression algorithm we find the best fit line based on the data that we have and this best fit line can be used for our predictions / decisions.

FINDING THE BEST FIT LINE:

After the best fit line is defined we will have 2 different values

1. Actual value as per the actual database = Y actual
2. Predicted value as per the best fit line.= Y predicted

In general the points on the best fit line is expected to have the least error from the actual values in the database.
So to find the best fit line a methodology called “Ordinary Least Square” method is followed where a term called “Residual Sum of Squares” is defined as
RSS = Sum of squares of (Y actual – Y predicted)

Based on the actual database the linear regression algorithm tries to fit in the best possible line which will have the least RSS.

The line with the least RSS will be declared as the best fit line for that particular model and we can compute the slope and intercept of this line. These inputs can then help to do any prediction / analysis or decision for that particular data model.

Linear Regression is categorised in 2 types.

1. Simple linear regression
   Contains only 1 predictor variable and the best fit line is defined as
   $Y = m X + c$

2. Multiple linear regression
   Contains multiple predictor variables and the best fit line is defined as
   $Y = m1 * X1 + m2 * X2 + m3 * X3 + \ldots\ldots + mn * Xn + c$

ASSUMPTIONS IN LINEAR REGRESSION :

Computation of linear regression is based on following assumptions

1. The error terms are normally distributed with peak at mean = 0
2. The error terms are not related
3. The error terms have a constant variance.

METRICS FOR MODEL BASED ON LINEAR REGRESSION :

The aim is that, out of the variance that is present in the target variable, maximum should be explained by the predictor variable.

R-squared is a metric which defines how much variance in the target variable is being explained by the other independent variables which a part of the linear regression model.

R-square = 1 – (RSS / TSS)

RSS : Residual sum of Squares

TSS: Total Sum of Squares.

R square value can vary between 0 and 1 and higher the value of R-square, more is the variance that is explained by the other variables and thus better is the model.

Generally R-square term is used for evaluation of simple linear regression model.

For multiple linear regression models another term Adjusted R-square is used, which also takes into account the total number of variables used in the multiple linear regression model to arrive a better metric for model evaluation.

Q. Explain the Anscombe's quartet in detail.

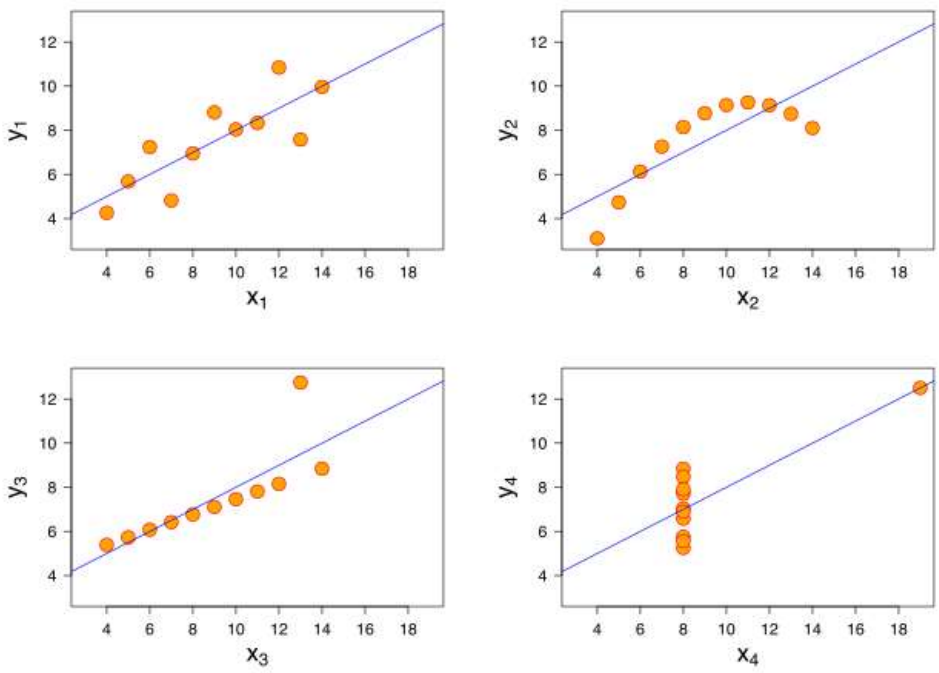
A. Whenever we are evaluating any dataset, no doubt we need to look at the statistical parameters of the dataset, however it is also necessary for us to visualise the dataset and Anscombe's quartet illustrates exactly why this is required.

Anscombe's quartet consist of 4 set of data containing X values and Dependent Y values. Each set has 11 values of X and Y.

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

On statistical evaluation of this dataset we find that the statistical values like mean, std. deviations correlation, linear regression line, are all same for all these datasets, and if we only evaluate these points we may come to a conclusion that all these 4 datasets are similar.

However when we plot a scatter plot for these datasets we get results something like these



From the above graphs we can see that all the datasets are actually visually very different and so only based on the statistical data we should not conclude that these datasets are same.

This quatret were found by James Anscombe with reasoning to prove that along with statistical analysis , it is also important to visualise the data to arrive at complete and proper results.

Q. What is Pearson's R

A. Corelation is defined as a statistical measure which describes a linear relationship between 2 variables.
Pearson's R is a coeffecient which quantitatively describes the corelation between 2 variables.
It is given by formula
Persons's R = sum [(Xi – Xmean) * (Yi – Ymean)] / sqrt{ [sum(Xi-Xmean)^2] * [sum (Yi – Ymean)^2]}
The values ranges between (-)1 to (+)1 with:
(-)1 indicating a perfect negetive linear relation
0 indicating no corelation
(+)1 indicating a perfect positive linear relation
The corelation increases as the corelation coefficient moves away from 0 either towards negetive or positive side.

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

A.
Scaling is a process wherein the values of the variables are brought within a similar scale. While building a model when the variables are on a very different scale, it is a good practice to bring the variables within a common scale which can help the model to learn the variables faster.

There are 2 types of scaling

1. Normalized scaling:
   Normalized scaling is done with below formula
   Xi – x(min) / x(max) – x(xmin)
   With this scaling, all the values of the variable are scaled within a range of 0 to 1.

2. Standardized scaling
   Standardized scaling is done with below formula
   Xi – mean(x) / standard deviation(x)
   With this scaling, the values of the variable will be such that their mean will be equal to 0 and the standard deviation will be 1.

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

A. VIF (Variance Inflation Factor) is a measure of multicolinearity between the predictor variables.
VIF is given by the formula 1 – (1/R-square)
For calculating VIF of any variable, first the R-square of the variable is calculated by studying the effect of other variables on the concerned variable. As we have seen in answer of an earlier question, the value of R-square varies from 0 to 1 and value of R-square as 1 indicates that the variance in target variable is perfectly explained by the other variables.

If we see the formula, if the value of R-square becomes 1 then VIF will turn out to be infinite thus indicating that the target variable is highly related (multicolinear) with the other variables, thus deeming it not useful for the model under construction.

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

A. Q-Q plot (Quantile – Quantile plot) plots the quantiles of Y set with respective quantile of X set.
A Q-Q plot is used to check if the values of the Y set follow the same distribution as that of the X set.
This is deduced by drawing a 45 degree straight line on the 2-D graph. If all the Q-Q plot points fall on this line it concludes that the values of Y set follows the same distribution as that of the X set.

For e.g. suppose we have a set whose distribution is unknown. We take a sample set of a known distribution for e.g. the sample set follows a normal distribution. Then when we plot a Q-Q plot of the unknown set with the sample set and if all the points fall on the 45 degree straight line then we can conclude that the unknown sample set also follows normal distribution.

This Q-Q plot can be used in linear regression to validate our assumption that the error terms are normally distributed.
Also if we have training and test dataset from different sources, then we can use Q-Q plot to check if both datasets are derived from the same population and both follow the same distribution.