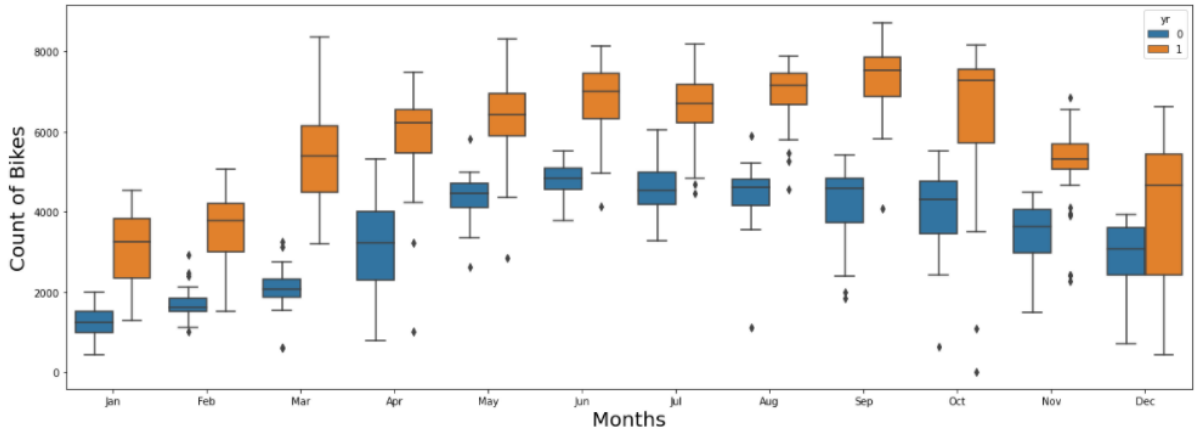
Assignment-based Subjective Questions

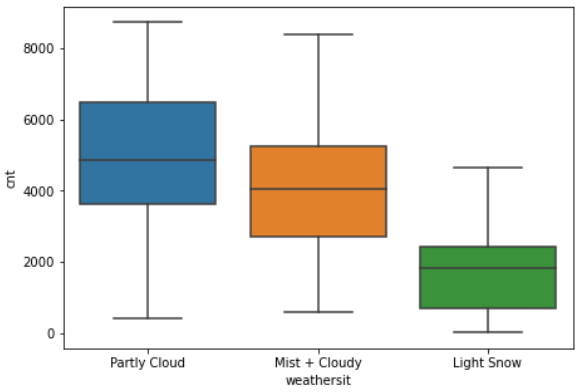
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

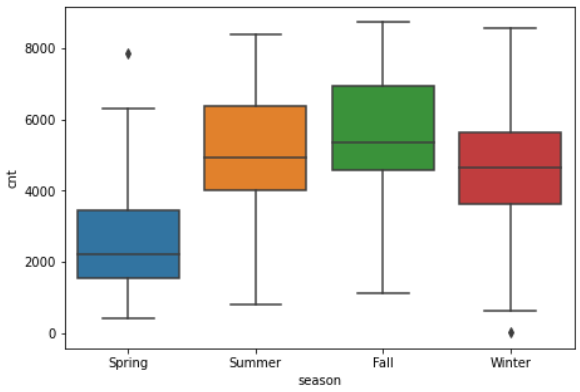
Answer: The categorical variables surely plays a good amount of role for predicting of dependent variables but visualizing the same gives us a fair amount of idea on which fields / category will be effective in building the model and getting the good accuracy. Below some of the point we could notice and involve them in our model building process.

- Month category we could see plays a fair amount of role, as from the month 5 to 10 we can see good amount of demand of bikes.
- Year is also giving a clear picture as the year progress forward the count or demand is getting higher.
- By combining both the fields we could, as the month and year progresses the demand is getting higher. PFB the snap of the same.



- Just like Month and Year, Season and Weather Situation categorical also plays good amount of role, as from the below snap we could see the median from Summer and Fall are close enough and also tells us the demand for bike increases in the same season.





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

Answer: The attributes used to drop the first column while encoding or creating the dummy variables. The thing which makes the attributes important is reducing the number of columns which gets created while creating the dummy variables. Reduction in the number of columns reduces the correlation among the variables too which in turns gives us the less chances of having Multicollinearity. For instance, if we have categorical variables of Gender which has a values like Male, Female and Other. Once we create a Dummy variable for the above by drop first let's consider we dropped "Other", then it becomes obvious if the values for Male and Female has 0 value, the person belongs to Other category.

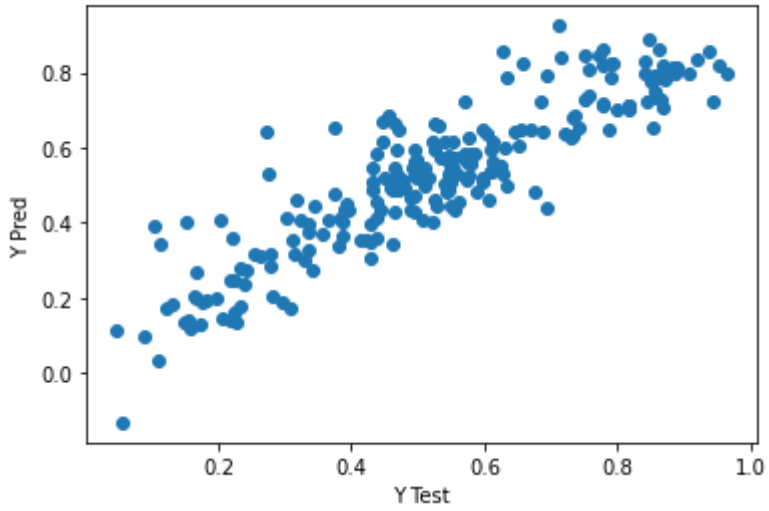
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: As when we checked by plotting the heatmap and the pair plot, the pair plot shows there is a linear relation between "temp" and "cnt" and the heatmap shows the "temp" variables has high correlation with "cnt".
(Note: there is another variable too "atemp" (Feel like temperature), but as we proceed along with building the model, we drop one of them. Besides both the variables having values close to close, so it could be the case where user gets to go the bike based on his feeling temperature.)

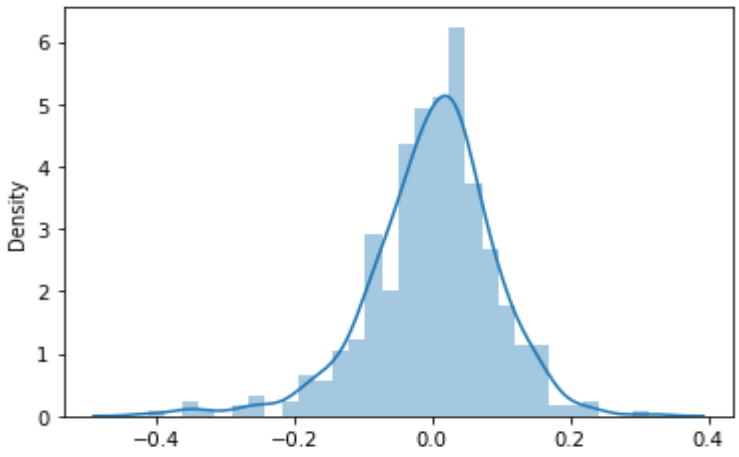
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: There are some assumptions for building the Linear Regression model. They are as follows :

- Linear relationship between X and y : There should be linear relationship between X and y, X being independent variable and y being dependent or predictor variable. If there is no linear relation, there is no use of fitting the model. Post making the model, we predict the value of y and plot the value with the y_test. "y_test" is nothing but predictor variable of test set. If the scatter graph so the linear form then our linear model was accurate.



- Error terms should have constant variance: The value of the error term should not increase or decrease as the error value changes.
- Error term or Residual should be normally distributed: The residual should be normally distributed. In most of the cases, it has been seen the residual follows the normal distribution with the mean equals to zero. Below refer the snap of residual distribution:



- Multicollinearity: The error terms should not be correlated with each other. They error terms should not dependent on each another.

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

Answer: Below are the three variables having high correlation:

- Temp (Positive correlation)
- Light Snow (Negative correlation)
- Year (Positive correlation)

General Subjective Questions

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

Answer: The Linear regression shows the linear relationship between the independent variable and dependent variable. The independent variables are those which are used to predict the dependent variable in short independent variable is the cause as on the other side, dependent variables are those which is to be predicted, we can say the effect. There are basically two types of linear regression. They are as follows:

- **Simple Linear Regression:** The linear regression which deals with single input variable (single independent variable) to predict, such linear regression is called Simple Linear Regression.
- **Multiple Linear Regression:** The linear regression which deals with multiple input variables (more than one independent variable) to predict the target/dependent variable, such linear regression is called Multiple Linear Regression.

The linear regression model gives a sloped straight line through the data points which describes the relation between the data point or variables.

As we know, to get the best fit line we use the slope-intercept form. The equation for the best straight line is

y = mx + b which is equivalent to

y = beta1x + beta0

where,

y = dependent variable

beta0 = intercept

beta1 = slope.

X = independent variable.

In Linear Regression, the Cost Function is used to find the accuracy of the mapping function which maps the input variables to output variable. The Cost Function is nothing but a function which helps us out to find the best possible values for beta0 and beta1, which in turn provides the best fit line for the data points. In Linear Regression, Mean Square Error cost function is used, which is the average of the squared error that occurred between predicted value and the actual value.

A regression can be Positive Linear Relationship or Negative Linear Relationship.

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

Answer: Anscombe's quartet can be defined as a group of four data set which are nearly identical in a simple statistic but there are minor differences which couldn't get highlighted unless we plot the dataset.

So, the idea behind Anscombe's is the importance of the visualizing the data.

Since we have four data set which are nearly identical as by just looking into those we wont be able to tell the difference, so by plotting or visualizing the data we will get to know the difference between the data set.

It was constructed by the statistician named Francis Anscombe to illustrate the importance of visualizing before proceeding for building the model or analysing. There are these four sets plots which have nearly same statistical observation like mean, variance for all x, y in all of the four data set.

This tells us the importance of visualizing the data before applying various algorithm to build the model out of them. Visualizing the data also helps us to see the distribution of the sample which can help us to see the wrong data present like outliers, missing value like NaN, NA etc, linearity among varibales etc.

Therefore, visualizing the data is very important step to understand and get the clarity on the data set and the factor which would helps us to build the model otherwise any regression algorithm can be fooled.

3. What is Pearson's R?

Answer: Pearson's correlation coefficient is the test of statistics that measures the statistical relationship between two continuous variables. It is known as the best method of measuring the correlation between the variables based on the method of covariance.

Is shows the linear relationship between two sets of data. The problem with this is, it is not able to tell the difference between dependent variables and independent variable. Along with this, it will not give you any information about the slope of the line, it will only tell us whether there is a relationship or not.

Pearson's correlation coefficient, when applied to a population, is normally represented by the Greek letter ***rho*** and may be referred as the population correlation coefficient or the population Pearson correlation coefficient.

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

Answer: Scaling of the variables is the part or step of preparing the data, scaling is applied on the independent variables to get them down to same level or more like normalizing the data within a particular range. It also helps in speeding up the calculation in an algorithm.

There are basically two types of scaling:

- Normalization / MinMax Scaling: It bring all the data in the range of 0 and 1. **MinMaxScaler** library from the **Scikit Learn Pre-processing** module helps to implement normalization in Python.
- Standardization Scaling: Standardization replaces the values by their Z scores. It bring all of the data into a standard normal distribution which has a mean zero and the standard deviation as one. Scale library from the same Scikit Lean Pre-processing package used to helps to implement the standardization in Python.

If the scaling is not done, then the algorithm only takes the magnitude in account and not the units which lead to incorrect building of model. Having said that, if the scaling is not done, you will have one column which has small values and one column which has large values so building the model with such values will not give proper number for the model, so the scaling needs to be done for such variable to bring them down to same level.

Although, it is important to note that scaling does not affects parameters like P-values, F-statistic, t-statistic et, it just affects the coefficients.

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

Answer: VIF is infinite which means the two independent variables has perfect strong correlation between them, which is nothing but multi-collinearity which needs to be removed to have accurate model. So, one of the way to deal with such issue is to drop one of the variables from the dataset, like in our case "temp" and "atemp" was having high correlation between them, when we build another model by dropping "atemp", we found out the multi-collinearity decreased drastically.

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

Answer: The Q-Q plot stands for Quantile – Quantile plot which is a graphical technique for determining if two data sets are coming from the population with a common distribution. It also helps us to assess if set of data came from theoretical distribution such as Normal distribution, Exponential distribution or Uniform distribution.

This helps in a 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 ate from the population with same distributions.

A Q-Q plot is also tells us about the shape of the distribution, location scale.

Below some of the interpretations for two data sets.

a) Similar distribution: If all the point of quantiles lien on or close to straight line at an angle of 45 degree from x-axis.
b) Y-values < X-values: if y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: if x-quantiles are lower than the y-quantiles.
d) Different distribution: If all the point of quantiles lies away from the straight line.

The package statsmodels.api provides qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.