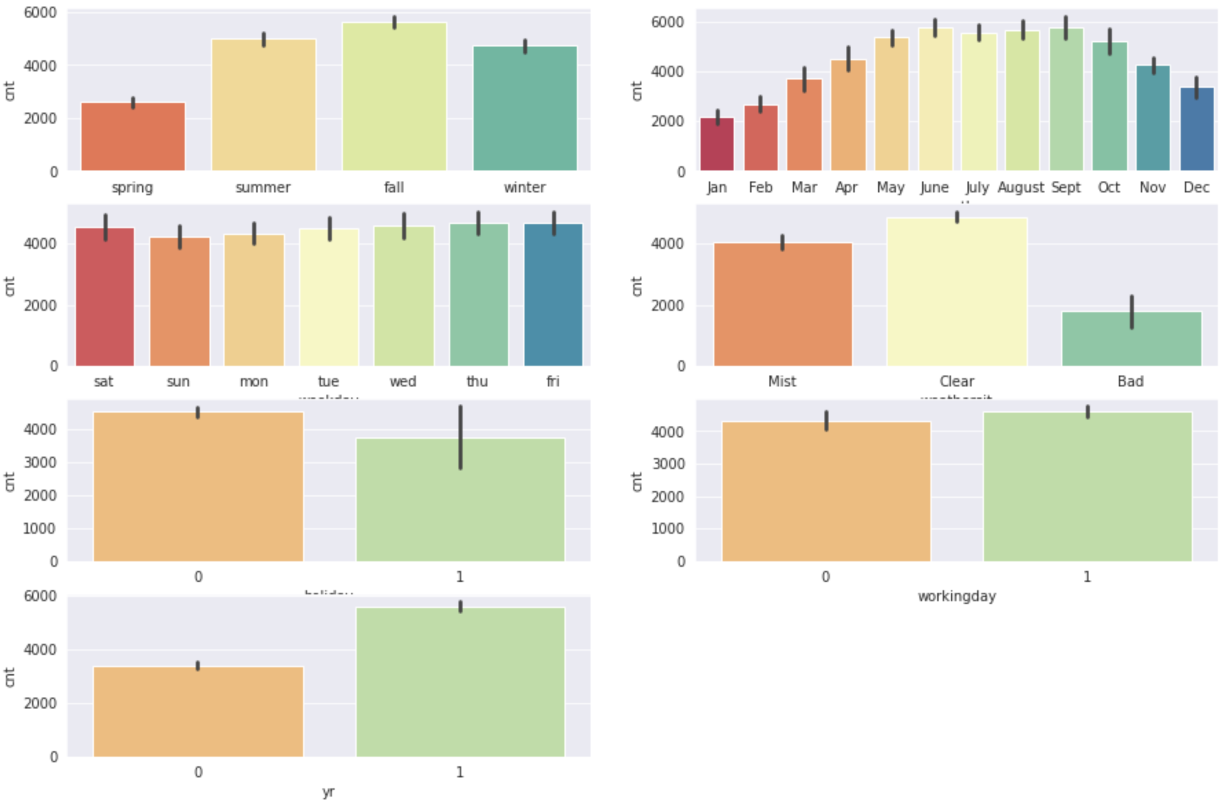
BIKE SHARING & LINEAR REGRESSION ASSIGNMENT

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

Ans- Following inferences are made after plotting the graphs of categorical variable.

1. Month wise during September bike sharing is more, compare to year ending and starting it is the lowest.
2. On holidays there is less demand of bike sharing.
3. Days of the week not clearly distinguish any clear data.
4. Bike sharing demand increases in clear weather conditions.
5. Bike demand for the next year is high.
6. Season wise Fall has the highest bike sharing demand.(1 denotes 2019)
7. Month-wise bike sharing demand is gradually increasing till September and it starts to decrease gradually.

Plots:



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

If we don’t use drop_first=True then the first column of the dummy variables table won’t be dropped. That would bring collinearity in the model and would make the model insignificant. Hence drop_first=True is important.

Also with n-1 dummy variables we can explain n category in a dataset.

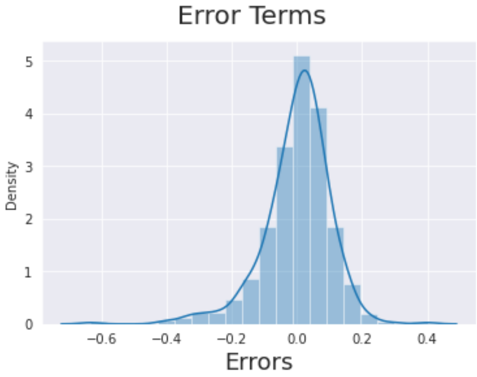
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The feeling temperature(atemp) has the highest correlation with the target variable cnt of value 0.63 which is similar to temperature.

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

Assumptions of linear Regression after building the model on training set is validated in following ways:

- **Error terms are normally distributed with mean zero.** We have plotted error terms in a distribution plot, plot result attached below which clearly explains that the error terms arenormally distributed with mean near to the zero.

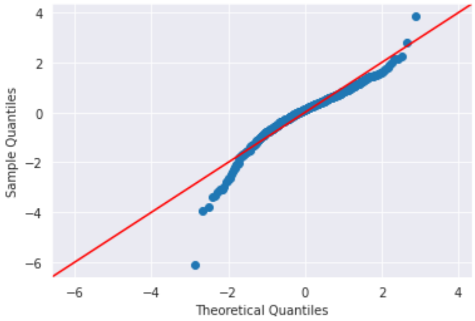


- **There should be no high multi collinearity among the features of the models.** To verify thecondition, we have calculated Variance Inflation Factors (VIF) for the final model features. VIF values are as follows, where all the VIF values are less than standard threshold value of 5.0.

```
            Features   VIF
2              atemp  4.90
1         workingday  4.12
3          windspeed  3.86
0                 yr  2.02
7        weekday_sat  1.71
4      season_summer  1.58
8   weathersit_Mist   1.50
5      season_winter  1.39
6          mnth_Sept  1.20
```

- Normal distribution on the residuals by plotting Q-Q plot.
  Plot shows the normal distribution of the residual’s values with

moderate fat tails.



-

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

Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are determined by the magnitude of the coefficients and they are :

1. atemp —> coef 0.58
2. Yr(2019)—> coef 0.24
3. Winter —> coef 0.10

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm shows a linear relationship between a dependent(y) variable and one or more independent variables(x). Linear regression helps build a model which helps us to find out how the value of the dependent variable is changing according to the value of the independent variables.

It helps us to do the following:

1. Finding out the effect of independent variables (x) on Target/ dependent variable (y)
2. Finding out the change in the target variable with respect to one or more input variable.
3. To predict out upcoming or the ongoing trends.

The equation for linear regression algorithm is:

$$y=b_0+b_1x+\text{ error}$$

The errors represent everything that the model does not have into account because it would be extremely unlikely for a model to perfectly predicta variable, as it is impossible to control every possible condition that may interfere with the response variable. The errors may also include reading or measuring inaccuracies as well.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet comprises of 4 data set that have nearly identical and simple descriptive analysis yet have very different distributions and appear very different when graphed. They have quite different distributions and appear differently when plotted on scatter plots. It fools the regression modelif built.

When the models particular to each of the datasets are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by the peculiarities.

Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to makea good model.

3. What is Pearson’s R?

- Pearson’s R is the test statistics that measures the statistical relationship, between two continuous variables. It gives information about the magnitude of the linear association, or correlation between 2variables. It is denoted by r.
- It also mentions whether there is a statistically significant relationship between any 2 variables.It also mentions about how 2 variables are strongly related to each other.
- Pearson coefficient is sensitive to outliers.

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

Scaling is the process of bringing all the features in the same standing since there might be some features whose units are very much different in magnitude than the rest of the features.

Scaling helps in making the model better. Due to difference in the units of the features, the correlation of the features takes a toll. 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. It does not give accurate results. Hence making the modelweak. Therefore, scaling is one of the most critical steps before creating a model.

The difference between Normalized Scaling and Standardized Scaling are as follows:

Normalized Scaling	Standardized Scaling
It scales and translates each feature individually such that it is in the given range on the training set between0 and 1.	It features and scales them suchthat the distribution centered around 0, with a standard deviation of 1.
If data has too many outliers then this method is not the best.	If data is not normally distributed, this is not the best method.

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

VIF is the measure of the extent of the correlation between one and other predictor variables in a model. It is used to check multi- collinearity. High values of VIF mean that there is high multicollinearity associated with the predictor variable.

An infinity value of VIF shows a perfect correlation between two predictor variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) which is infinity.

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

Q-Q plot stands for Quantile-Quantile plot which basically plots the quantiles of first data set with the quantiles of the second data set.

The uses and benefits of this plot are:

- It can detect outliers.
- Change in scale, symmetry can be detected.
- Can be used with any sample of data.

The importance of Q-Q plots is that they are used to assess whether a variableis normal or not. We can use Q-Q plots to check our data against any distribution, not just the normal distribution. Since the methods we apply are mostly based on normality assumptions, it is important to check the normalityof the sample data. This is where Q-Q plots come into the picture.