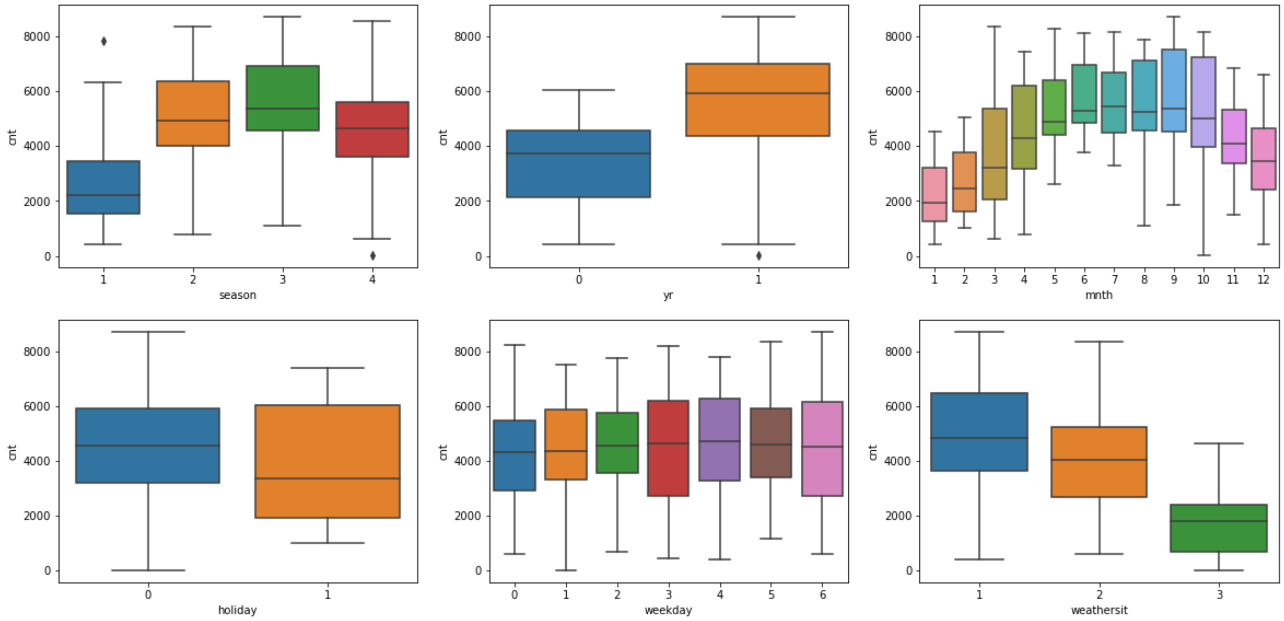
Assignment-based Subjective Questions

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

We have 6 categorical variable in our data set and by looking at the box plot we can infer its affect on bike uasge count as follows.



1) season : season (1:spring, 2:summer, 3:fall, 4:winter) :there is more usage of bike in summer and fall compared to spring and winter.

2) yr : year (0: 2018, 1:2019) : In the year 2019 we see more usage of bike than in the year 2018

3)mnth : month (1 to 12) : Usage of bike is more in the middle of months like from 3 to 10 we see more usage and this what we can map with the seson column as well.

4)holiday : weather day is a holiday or not : we dont see how of differenec in the pattern of bike usage if its holiday or not.

5) weekday : day of the week : we use same usage pattern for all the weekdays.

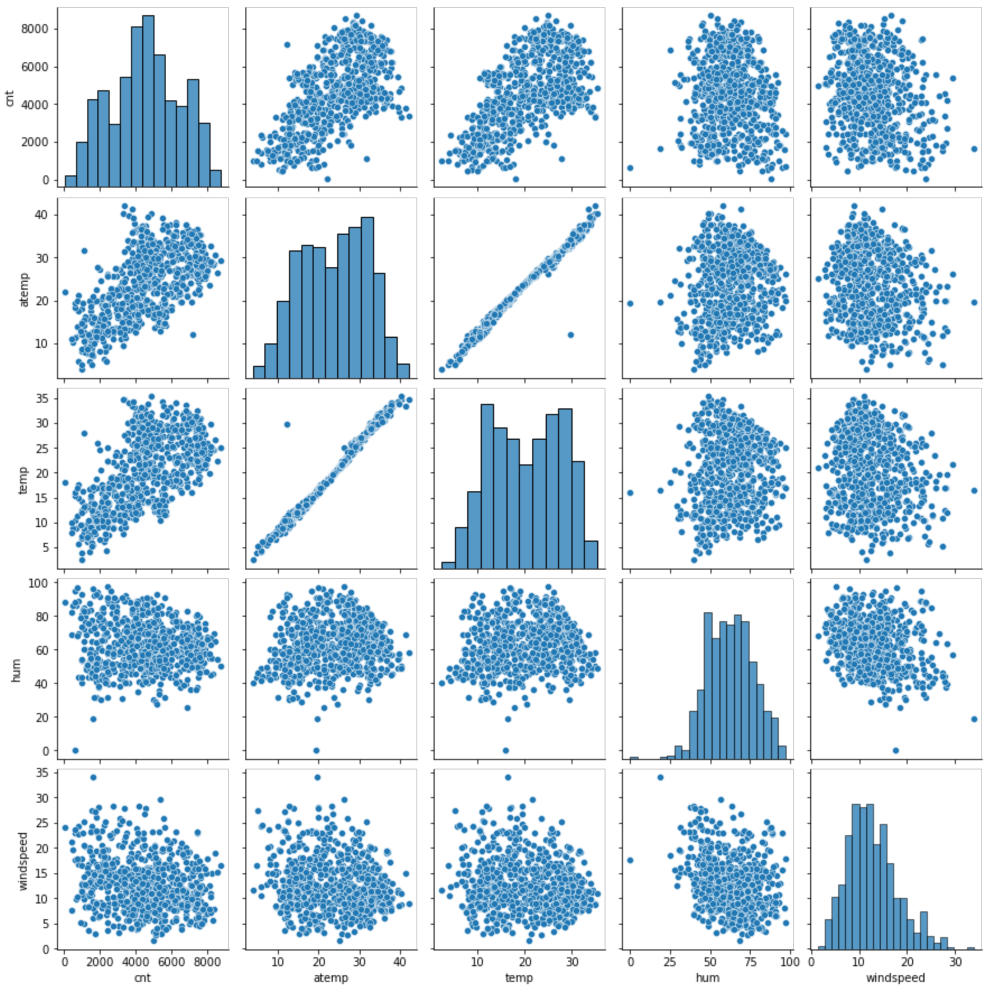
6)weathersit : Clear,Mist + Cloudy,Light Snow,Heavy Rain : More usage of bike when weather is clear.

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

WhenWhen we generate a dummy variables of the categorical column with K level, we end up with K categorical variables and there will be one redendency of information, so we use drop_first=true to delete the reference column after generating the dummies.

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

temp and atemp has highest correlation with cnt variable.



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

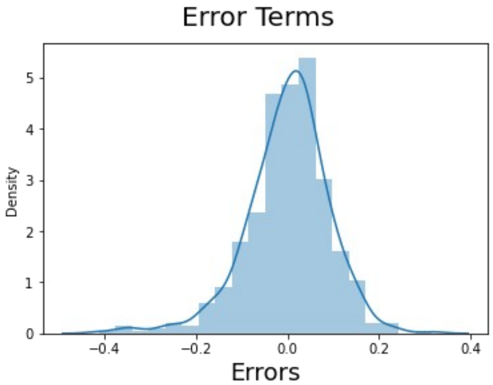
Generate VIF after every iteration and drop the columns with a VIF value > 5 in the below order(and p value > 0.05) and check the error residual distribution and It should be normally distributed.

1) High P Value and high VIF value

2) High P value and low VIF

3) Low p value and high VIF

4) Low p value and low VIF



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

1) temp
2) season (2 and 4)
3) weekday(6)

General Subjective Questions

Explain the linear regression algorithm in detail.
Linear regression is the supervised Machine Learning model and its a process of estimating the relationship among the variables and to establish the relationship between the dependent and independent variables by fitting the best fit linear line with minimized error. Its used for forcasting and prediction on the data which follows below list of assumption.

1) linearity : There should be linear relationship between dependent and independent variable, for ever unit change independent variable there will be constant amout of change.

Equation of LR : y = b0x0 +b1x1+.......bnxn

2) Error terms are normally distributed.

3) Error having constant varience

4) Homostedicity

5) No Autoco-relation

Explain the Anscombe's quartet in detail.

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.

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**.**

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

What is Pearson’s R?

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

Scaling is to standerdizing the value of all the numerical variable.
When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with different scale of coefficients that might be difficult to interpret.
So we need to scale features because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods
There are two ways to scale :
1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
Its because of strong correlation between indepenedent variable and it shows multicorrelation.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of 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 are from populations with same distributions.