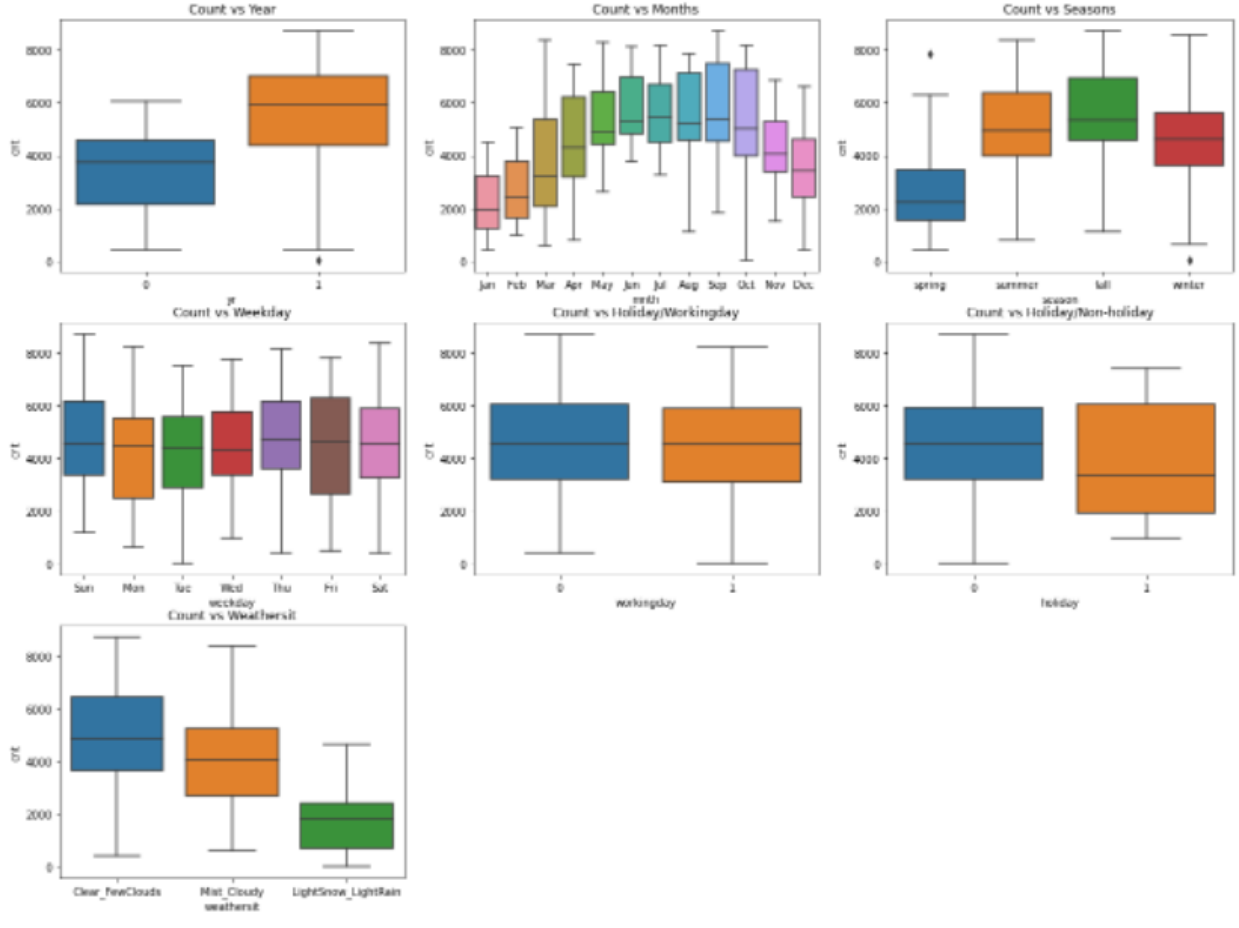
Assignment-based Subjective Questions with Answers:

Q1. ***From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?***

A. The demand of bike sharing got higher in the 'year' 2019 as compared to the 'year' 2018. Also, during the 'month of August' and 'month of September', when the 'temperature' is neither too hot nor too cold, there was a hike in demand of bike sharing. Which indicates, 'Seasons' affect the bike sharing demand market as well. However, during the 'holidays' there was not much of a noticeable demand, the reason might being people willing to spend time with family. The important categorical features affecting the demand of bike sharing are – Year, Temperature, Months and Seasons. Sharing the box plot from data visualization.



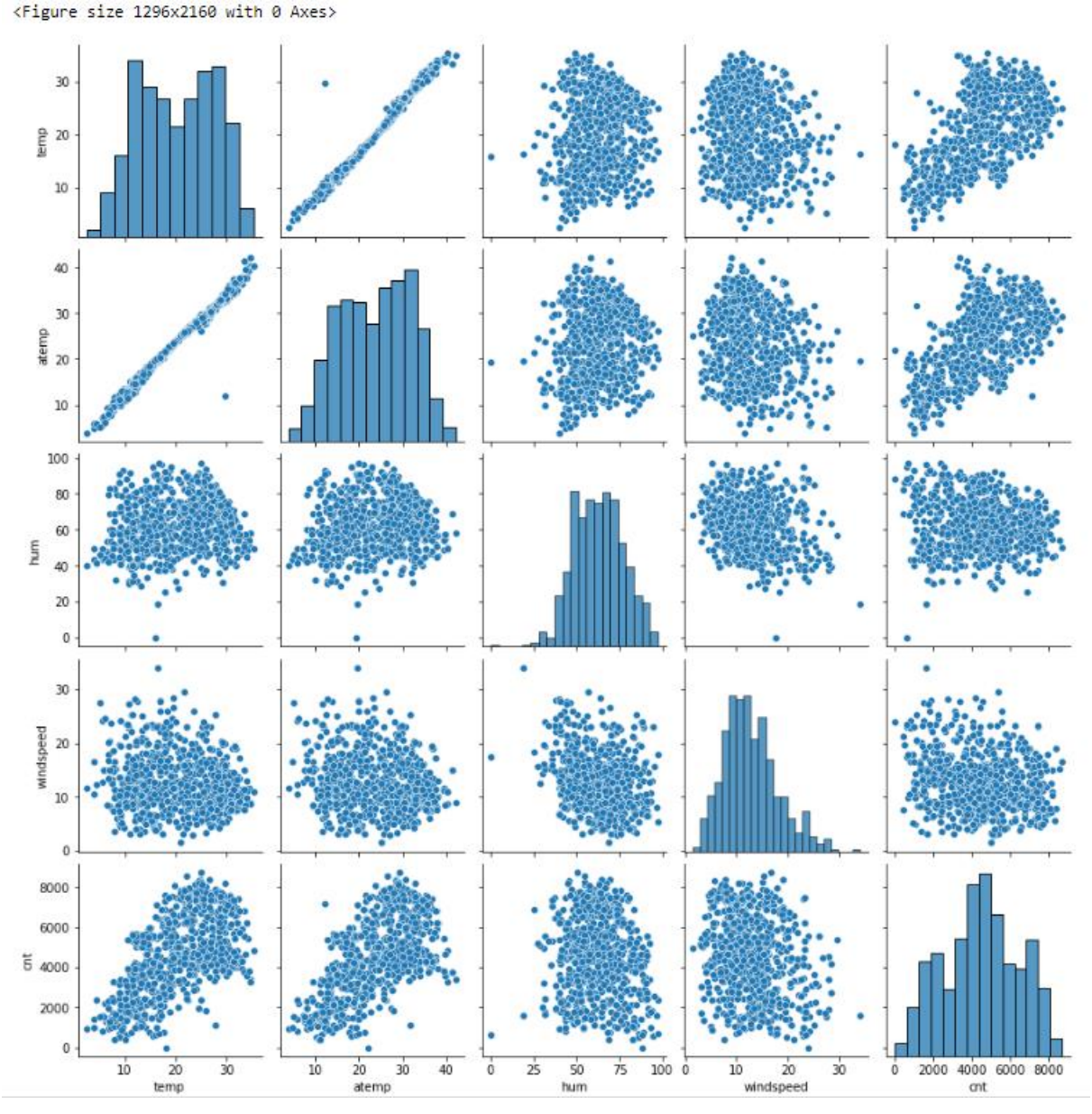
Q2. Why is it important to use drop_first=True during dummy variable creation?

A. During dummy variable creation, we first look at the categorical variable and understand how many levels does it have. For example, while we are creating dummy variables for column – 'season' we know, we have 4 seasons in total which means we have 4 level of categories here. Spring, summer, fall and winter. But we have a rule to follow while creating dummy variables. That is, if we have K number of levels in a category, then we need to create (K-1) dummy variables. We can do it manually or let python do it for us.

In order to do it using python, we use “drop_first=True”. As, python will create K levels of dummy variables, and “drop_first=True” will help in removing that one extra column and maintain the need of (K-1) variables. Hence, it reduces the correlations created among dummy variables.

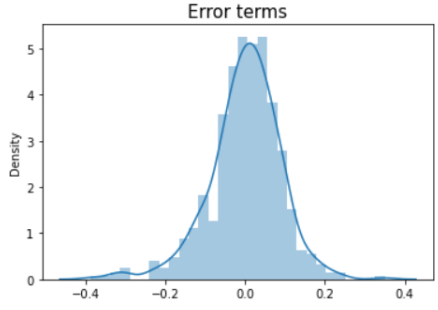
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A. Looking at the pair-plot among the numerical variables, we can say that our target variable ‘cnt’ has the highest correlation with independent variables ‘temp’ (temperature) and ‘yr’ (year) for sure. ‘temp’ and ‘atemp’ seems to be having strong correlation as well, which may lead to multicollinearity. So, we had to drop ‘atemp’.



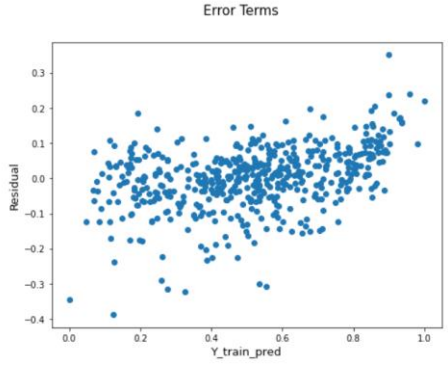
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A. After building the model, we evaluate it before testing the model with Test dataset. And to do so, we plot a distribution and scatter plot to check the error terms.
In our case, we could evidently see that in the distribution plot the mean centered on zero and in the scatter plot the error terms were completely random suggesting that it was a normal distribution. Also, from the scatter plot, we could see the error terms to be having constant variance which is also known as homoscedasticity.



Insight:

- We can see the mean of the plot is around zero and it looks like a normal distribution



Insight:

- The error terms, looks like, to be having constant variance which is also known as homoscedasticity.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A. Based on the final model, we can conclude that the 3 most important features having significant contribution towards the demand of the shared bikes are: Temperature, year, and mnth_sep. During pleasant temperature in the year of 2019 and in the month of September the demand was in peak. So, it advisable to prioritize these variables while planning to achieve maximum demand.
Temp – 0.540
Year – 0.229
Mnth_sep – 0.112

General Subjective Questions and Answers:

Q1. Explain the linear regression algorithm in detail.

A. Linear Regression is a statistical method used for predictive analysis for numeric variables where we'll have one target variable against one or more independent variable. Our sole purpose of using Linear Regression is getting the 'Best Fit Line' representing the relationship between variables.

Algorithm I have followed:-

- Read and Understand the data:-
    - Import necessary libraries, check info, shape and statistical description of dataset
- Clean the dataset (if needed):-
    - Convert the data type if needed, check for null and duplicate values
    - Drop the columns which are unnecessary
- Visualizing the data:-
    - Pair plot among numerical variables
    - Box plot among categorical variables
    - Heat map for an idea of the correlation again the target variable
- Data Preparation:-
    - Create dummy variables for the categorical features with more than two levels
- Splitting the data into Training and Testing dataset and Scaling:-
    - Split the entire data set in 70:30 ratio for train and test respectively
    - Also, scale all the features using MinMaxScaler and compress the range between 0 and 1
- Building and Training Linear Models:-
    - Assign y_train and X_train
    - Perform RFE to only keep the most important 15 features
    - Add coefficient constant to train data
    - Build the model with significant p-values and VIF
- Residual Analysis and Evaluation of the Train Data:-
    - Calculate the residual
    - Plot error terms on distribution plot to check if it is a normal distribution or not
    - Plot error terms on scatter plot to check the constant variance
- Making Predictions using the Final model:-
    - Scale the test data as we did before
    - Assign y_test and X-test using final model features
- Residual Analysis and Evaluation of the Test Data
    - Calculate residual
    - Calculate r-squared, adj. r-squared, mean squared error
    - Compare them with our train data values
    - Plot the error terms on distribution plot and scatter plot
    - Also check y_test vs. y_pred on a scatter plot
- Conclusion:-

- Conclude your findings and recommendations.

Q2. Explain the Anscombe's quartet in detail.

A. Anscombe's quartet is made up of four datasets with very similar statistical description including variance, mean and X, y points, but when you graph them they look very different from each other, also, have very different distributions.
Statistician Francis Anscombe has derived this while building a model and emphasized on plotting graphs before building a model as the anomalies like outliers, and linear productivity factor of data present in the datasets can fool the regression model.
If we plot visualization on these datasets we can check if they fit they model or not, if there are outliers present which cannot be handled by regression model.

Q3. What is Pearson's R?

A. Pearson's R, also known as Pearson Correlation Co-efficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation. Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation
For example, we would normally expect the salary and tax to have a Pearson's correlation coefficient significantly greater than 0 but less than 1 (as 1 would represent an unrealistically perfect correlation).

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A. Most of the times, the data in our data set has features which highly vary in magnitudes, units and range. If scaling is not performed before building the model then algorithm only takes magnitude in account and not the units. Hence, we end up with incorrect modelling and predictions. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
For example, in our dataset temp has different unit and humidity, wind speed and target variable 'cnt' have different units. So, if we miss out on scaling these features the model would only consider the magnitude of the data and not the unit.

Normalization/Normalized Scaling – Also known as Min Max scaling – Compresses the entire dataset within the range of 0 and 1. Formula – MinMaxScaling:X= (X-Xmin)/(Xmax-Xmin).

Standardization – In this process, we replace the values with their Z-score. Hence, it brings all the data into a standard normal distribution which has a mean value zero and standard deviation (sigma).
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A. If there is perfect correlation among the features, then VIF is infinity. This indicates a perfect correlation between two independent variables. In this case of perfect correlation, we get R2 =1, which lead to VIF = 1/(1-R2) = infinity. To get rid of this situation we need to drop one of the variables causing perfect multicollinearity. For example – In our dataset, we had two feature 'temp' and 'atemp' they had a strong and perfect correlation which would have caused an infinity VIF value. So, to avoid that multicollinearity we dropped 'atemp'.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A. Q-Q Plots or Quantile-Quantile plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. (median = 50% data below, 50% data above). The sole intention of Q-Q plots is to find out if two sets of data come from the same distribution. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale etc are similar or different in the two distributions.