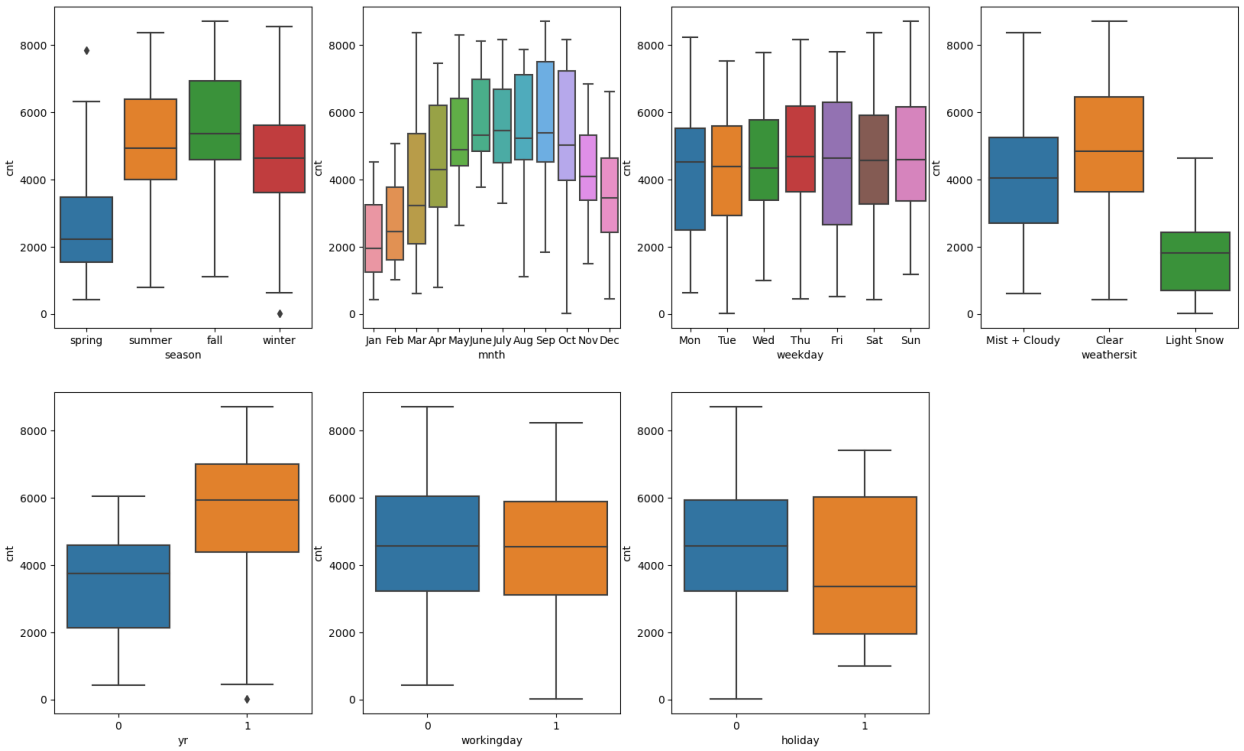
Assignment-based Subjective Questions

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



- Number of people renting bikes is more in summer and fall seasons and less in spring season.
- We see the bike renting is more in September and October months and less in January and February.
- The weekday boxplot is not of a great significance.
- More people rent bikes on a clear weather day.
- Number of bikes rented was more in 2019 compared to 2018.
- There is no significant difference in the number of bikes rented on weekday and weekend.
- More bikes were rented on a holiday.

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

drop first=True means dropping the first column**.** The goal is to reduce the number of columns by dropping the column that is not necessary. If you don't drop the first column, then your dummy variables will be correlated (redundant). This may affect some models adversely as there is redundancy.

For example: Suppose we have a gender column that has 3 variables male, female, other. So, a person can either be male or female and if they are not either of these, then their gender is other. Instead of having all the 3 variables we can just have 2. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

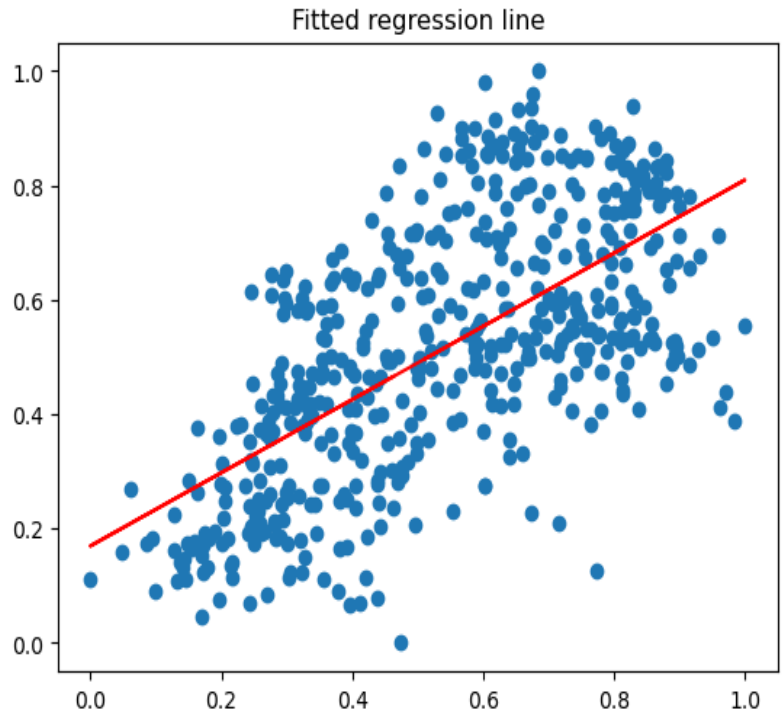
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   temp and atemp have highest correlation with the target variable.

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

   Validating the assumption of Linear Regression Model:

   - Linear Relationship



   The above graph represents the correlation between count and temperature with the regression line. We see that the linearity is preserved.

   - Autocorrelation in residuals(Independence of residuals)

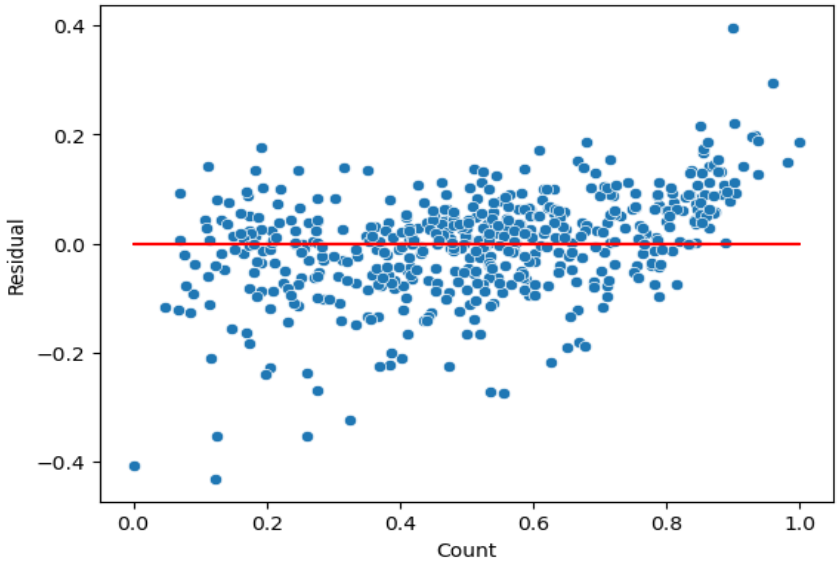
Omnibus:	66.905	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	173.340
Skew:	-0.661	**Prob(JB):**	2.29e-38
Kurtosis:	5.532	**Cond. No.**	13.2

Autocorrelation refers to the fact that observations' errors are correlated. To verify that the observations are not auto-correlated, we can use the Durbin-Watson test.
The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables.

0 – 2: positive auto-correlation
2 – 4: negative auto-correlation)

The Durbin-Watson value is 2.010 hence the independence of residuals is satisf

- Homoscedasticity



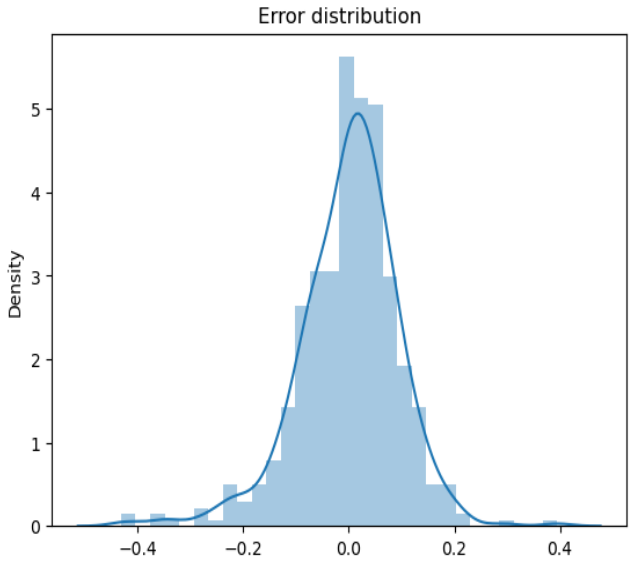
There is no visible pattern in residual values, thus homoscedacity is well preserve

- Absence of Multicollinearity

	Features	VIF
2	temp	2.99
0	yr	2.05
6	Mist + Cloudy	1.51
3	July	1.33
8	winter	1.33
7	spring	1.25
4	Sep	1.19
5	Light Snow	1.06
1	holiday	1.04

The VIF values of all the variables in less than 5 which shows absence of multicolinearity.

- Normality of Errors



The error terms is normally distributed around 0 as expected.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- Temp
- Weathersit
- yr

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning method that is used find a linear equation that best describes the correlation of the independent variables with the dependent variable.

- This is achieved by fitting a line to the data using least squares.

- The residual is the distance between the line and the actual value of the explanatory variable.
- The line tries to minimize the sum of the squares of the residuals.
- Finding the line of best fit is an iterative process.

Linear regression equation:

y=b0+b1X1+b2X2+……

where y is the dependent variable. **X1, X2, X3** etc. are the independent variables **b0, b1, b2** etc. are the correlation coefficients that explain the correlation between the dependent and the independent variables.

The sign of the coefficients determines if the variables is positively or negatively correlated.

b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

- A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is.
- The terms like R-squared values and p-value are used in linear regression.
- The R-squared value indicates how much of the variation in the dependent variable can be explained by the independent variables.
- The R-squared values range between 0 and 1. The nearer it is to 1 ,the better is the variation explained by the independent variable.
- The p-value explains how reliable that explanation is.
- A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model.

2) 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 they have very different distributions and appear differently when plotted on scatter plots.

Importance of Anscombe's quartet:
It tells us about the importance of visualizing data before applying various algorithms to build models.

3) What is Pearson’s R?

Pearson's r is the strength of the linear association between the variables.

- When r is between 0 and 1: there is positive correlation. When one variable changes, the other variable changes in the same direction**.**

- When r is 0: there is no correlation. There is no relationship between the variables.
- When r is between -1 and 0: There is negative correlation. When one variable changes, the other variable changes in the opposite direction**.**

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range i.e to rescale the values within the range [0,1]. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence result in incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized Scaling	**Standardized Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
SkLearn provides a transformer called MinMaxScaler for Normalization.	SkLearn provides a transformer called StandardScaler for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

The formula of VIF is given by:

1/(1-R(squared))

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential.

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

Uses of QQ plot:

- If two populations are of the same distribution.
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution